

# Probabilistic Models in the Study of Language

Roger Levy

November 30, 2010

# Contents

|          |  |           |
|----------|--|-----------|
| <b>2</b> | <b>Univariate Probability</b>  | <b>6</b>  |
| 2.1      | What are probabilities, and what do they have to do with language? . . . . .             | 6         |
| 2.2      | Sample Spaces . . . . .  | 8         |
| 2.3      | Events and probability spaces . . . . .  | 8         |
| 2.4      | Conditional Probability, Bayes' rule, and Independence . . . . .                         | 9         |
| 2.4.1    | Bayes' rule . . . . .  | 10        |
| 2.4.2    | (Conditional) Independence . . . . .   | 12        |
| 2.5      | Discrete random variables and probability mass functions . . . . .                       | 13        |
| 2.5.1    | Bernoulli trials and the Bernoulli distribution . . . . .                                | 14        |
| 2.5.2    | Multinomial trials . . . . .   | 14        |
| 2.6      | Cumulative distribution functions . . . . .  | 15        |
| 2.7      | Continuous random variables and probability density functions . . . . .                  | 16        |
| 2.7.1    | The uniform distribution . . . . .   | 17        |
| 2.7.2    | Cumulative distribution functions for continuous random variables . .                    | 18        |
| 2.8      | Normalized and unnormalized probability distributions . . . . .                          | 19        |
| 2.9      | Expected values and variance . . . . .   | 21        |
| 2.9.1    | Expected value . . . . .   | 21        |
| 2.9.2    | Variance . . . . .   | 21        |
| 2.10     | The normal distribution . . . . .  | 22        |
| 2.10.1   | Standard normal random variables . . . . .   | 23        |
| 2.11     | Estimating probability densities . . . . .   | 23        |
| 2.11.1   | Discrete probability densities: relative-frequency estimation . . . . .                  | 24        |
| 2.11.2   | Estimating continuous densities: histograms and kernel density estimation                | 25        |
| 2.11.3   | Kernel estimation for discrete densities . . . . .                                       | 27        |
| 2.11.4   | Kernel density estimation and exemplar models . . . . .                                  | 28        |
| 2.11.5   | Evaluating estimated probability densities . . . . .                                     | 29        |
| 2.12     | Exercises . . . . .  | 31        |
| <b>3</b> | <b>Multivariate Probability</b>  | <b>35</b> |
| 3.1      | Joint probability mass and density functions . . . . .                                   | 35        |
| 3.1.1    | Joint cumulative distribution functions . . . . .  | 36        |
| 3.2      | Marginalization . . . . .  | 36        |
| 3.3      | Linearity of expectation, covariance, correlation, and variance sums of random variables | 38        |

|          |  |           |
|----------|--|-----------|
| 3.3.1    | Linearity of the expectation . . . . .   | 38        |
| 3.3.2    | Covariance . . . . .   | 38        |
| 3.3.3    | Covariance and scaling random variables . . . . .                              | 39        |
| 3.3.4    | Correlation . . . . .  | 39        |
| 3.3.5    | Variance of the sum of random variables . . . . .                              | 39        |
| 3.4      | The binomial distribution . . . . .  | 40        |
| 3.4.1    | The multinomial distribution . . . . .   | 40        |
| 3.5      | Multivariate normal distributions . . . . .                                    | 41        |
| 3.6      | The central limit theorem . . . . .  | 42        |
| 3.7      | Exercises . . . . .  | 43        |
| <b>4</b> | <b>Parameter Estimation</b>  | <b>46</b> |
| 4.1      | Introduction . . . . .   | 46        |
| 4.2      | Desirable properties for estimators . . . . .                                  | 47        |
| 4.2.1    | Consistency . . . . .  | 48        |
| 4.2.2    | Bias . . . . .   | 48        |
| 4.2.3    | Variance (and efficiency) . . . . .  | 48        |
| 4.3      | Frequentist parameter estimation and prediction . . . . .                      | 49        |
| 4.3.1    | Maximum Likelihood Estimation . . . . .  | 50        |
| 4.3.2    | Limitations of the MLE: variance . . . . .                                     | 50        |
| 4.3.3    | Limitations of the MLE: bias . . . . .   | 51        |
| 4.4      | Bayesian parameter estimation and density estimation . . . . .                 | 55        |
| 4.4.1    | Anatomy of inference in a simple Bayesian model . . . . .                      | 55        |
| 4.4.2    | The beta distribution . . . . .  | 56        |
| 4.4.3    | Simple example of Bayesian estimation with the binomial distribution . . . . . | 57        |
| 4.5      | Computing approximate Bayesian inferences with sampling techniques . . . . .   | 61        |
| 4.6      | Further reading . . . . .  | 70        |
| 4.7      | Exercises . . . . .  | 70        |
| <b>5</b> | <b>Confidence Intervals and Hypothesis Testing</b>                             | <b>72</b> |
| 5.1      | Bayesian confidence intervals . . . . .  | 72        |
| 5.2      | Bayesian hypothesis testing . . . . .  | 73        |
| 5.2.1    | More complex hypotheses . . . . .  | 76        |
| 5.2.2    | Bayes factor . . . . .   | 77        |
| 5.2.3    | Example: Learning contextual contingencies in sequences . . . . .              | 78        |
| 5.3      | Frequentist confidence intervals . . . . .                                     | 79        |
| 5.4      | Frequentist hypothesis testing . . . . .                                       | 80        |
| 5.4.1    | Hypothesis testing: binomial ordering preference . . . . .                     | 82        |
| 5.4.2    | Quantifying strength of association in categorical variables . . . . .         | 84        |
| 5.4.3    | Testing significance of association in categorical variables . . . . .         | 86        |
| 5.4.4    | Likelihood ratio test . . . . .  | 89        |
| 5.5      | Exercises . . . . .  | 90        |

|          |   |           |
|----------|---|-----------|
| <b>6</b> | <b>Generalized Linear Models</b>  | <b>97</b> |
| 6.1      | The form of the generalized linear model . . . . .                          | 97        |
| 6.2      | Linear models and linear regression . . . . .                               | 98        |
| 6.2.1    | Fitting a linear model . . . . .  | 99        |
| 6.2.2    | Fitting a linear model: case study . . . . .                                | 100       |
| 6.2.3    | Conceptual underpinnings of best linear fit . . . . .                       | 100       |
| 6.3      | Handling multiple predictors . . . . .                                      | 102       |
| 6.3.1    | Residualizing . . . . .   | 102       |
| 6.3.2    | Multiple linear regression . . . . .  | 104       |
| 6.4      | Confidence intervals and hypothesis testing for linear regression . . . . . | 104       |
| 6.5      | Hypothesis testing in multiple linear regression . . . . .                  | 107       |
| 6.5.1    | Decomposition of variance . . . . .   | 108       |
| 6.5.2    | Comparing linear models: the $F$ test statistic . . . . .                   | 108       |
| 6.5.3    | Model comparison: case study . . . . .                                      | 110       |
| 6.6      | Analysis of Variance . . . . .  | 111       |
| 6.6.1    | Dummy variables . . . . .   | 111       |
| 6.6.2    | Analysis of variance as model comparison . . . . .                          | 113       |
| 6.6.3    | Testing for interactions . . . . .  | 114       |
| 6.6.4    | Repeated Measures ANOVA and Error Stratification . . . . .                  | 115       |
| 6.6.5    | Condition-specific random effects and error stratification . . . . .        | 118       |
| 6.6.6    | A case study . . . . .  | 120       |
| 6.7      | Other generalized linear models . . . . .                                   | 127       |
| 6.7.1    | Logit models . . . . .  | 127       |
| 6.7.2    | Fitting a simple logistic regression model . . . . .                        | 128       |
| 6.7.3    | Multiple logistic regression . . . . .                                      | 130       |
| 6.7.4    | Transforming predictor variables . . . . .                                  | 132       |
| 6.7.5    | Multiplicativity of the odds . . . . .                                      | 132       |
| 6.8      | Confidence intervals and model comparison in logit models . . . . .         | 133       |
| 6.8.1    | Frequentist Confidence intervals for logit models . . . . .                 | 133       |
| 6.8.2    | Bayesian confidence intervals for logit models . . . . .                    | 134       |
| 6.8.3    | Model comparison . . . . .  | 134       |
| 6.8.4    | Dealing with symmetric outcomes . . . . .                                   | 135       |
| 6.9      | Log-linear and multinomial logit models . . . . .                           | 137       |
| 6.10     | Log-linear models of phonotactics . . . . .                                 | 138       |
| 6.10.1   | Log-linear models . . . . .   | 142       |
| 6.10.2   | Translating between logit models and log-linear models . . . . .            | 151       |
| 6.11     | Guide to different kinds of log-linear models . . . . .                     | 152       |
| 6.12     | Feed-forward neural networks . . . . .                                      | 154       |
| 6.13     | Further reading . . . . .   | 154       |
| 6.14     | Notes and references . . . . .  | 155       |
| 6.15     | Exercises . . . . .   | 155       |

|           |   |            |
|-----------|---|------------|
| <b>8</b>  | <b>Hierarchical Models</b>  | <b>162</b> |
| 8.1       | Introduction . . . . .  | 163        |
| 8.2       | Parameter estimation in hierarchical models . . . . .                     | 165        |
| 8.2.1     | Point estimation based on maximum likelihood . . . . .                    | 167        |
| 8.2.2     | Bayesian posterior inference in hierarchical models . . . . .             | 168        |
| 8.2.3     | Multivariate responses . . . . .  | 170        |
| 8.3       | Hierarchical linear models . . . . .                                      | 171        |
| 8.3.1     | Hypothesis testing in hierarchical linear models . . . . .                | 174        |
| 8.3.2     | Heteroscedasticity across clusters . . . . .                              | 177        |
| 8.3.3     | Multiple clusters per observation . . . . .                               | 178        |
| 8.4       | Hierarchical generalized linear models . . . . .                          | 181        |
| 8.4.1     | Hierarchical logit models . . . . .                                       | 182        |
| 8.4.2     | Fitting and interpreting hierarchical logit models . . . . .              | 182        |
| 8.4.3     | An example . . . . .  | 182        |
| 8.4.4     | Model comparison & hypothesis testing . . . . .                           | 186        |
| 8.4.5     | Assessing the overall performance of a hierarchical logit model . . . . . | 186        |
| 8.5       | Further Reading . . . . .   | 187        |
| 8.6       | Exercises . . . . .   | 188        |
| <b>9</b>  | <b>Dimensionality Reduction and Latent Variable Models</b>                | <b>192</b> |
| 9.1       | Gaussian mixture models . . . . .   | 192        |
| 9.1.1     | Inference for Gaussian mixture models . . . . .                           | 196        |
| 9.1.2     | Learning mixture probabilities . . . . .                                  | 200        |
| 9.2       | Latent Dirichlet Allocation and Topic Models . . . . .                    | 202        |
| 9.2.1     | Formal specification . . . . .  | 204        |
| <b>10</b> | <b>Probabilistic Grammars</b>   | <b>208</b> |
| 10.1      | Outline . . . . .   | 208        |
| 10.2      | Weighted Finite-State Transducers . . . . .                               | 208        |
| 10.3      | Probabilistic context-free grammars . . . . .                             | 209        |
| 10.3.1    | Probabilities from wCFGs . . . . .  | 210        |
| 10.4      | Exercises . . . . .   | 211        |
| <b>A</b>  | <b>Mathematics notation and review</b>                                    | <b>213</b> |
| A.1       | Sets ( $\{\}, \cup, \cap, \emptyset$ ) . . . . .                          | 213        |
| A.1.1     | Countability of sets . . . . .  | 213        |
| A.2       | Summation ( $\sum$ ) . . . . .  | 213        |
| A.3       | Product of a sequence ( $\prod$ ) . . . . .                               | 214        |
| A.4       | ”Cases” notation ( $\{$ ) . . . . .                                       | 214        |
| A.5       | Integration ( $\int$ ) . . . . .  | 215        |
| A.5.1     | Analytic integration tricks . . . . .                                     | 216        |
| A.5.2     | Numeric integration . . . . .   | 218        |
| A.6       | Precedence ( $\prec$ ) . . . . .  | 218        |

|          |   |            |
|----------|---|------------|
| A.7      | Combinatorics ( $\binom{n}{r}$ ) . . . . .                                | 220        |
| A.8      | Basic matrix algebra . . . . .  | 220        |
| A.9      | Miscellaneous notation . . . . .  | 222        |
| <b>B</b> | <b>More probability distributions and related mathematical constructs</b> | <b>223</b> |
| B.1      | The gamma and beta functions . . . . .                                    | 223        |
| B.2      | The Poisson distribution . . . . .  | 224        |
| B.3      | The hypergeometric distribution . . . . .                                 | 224        |
| B.4      | The chi-square distribution . . . . .                                     | 224        |
| B.5      | The $t$ -distribution . . . . .   | 225        |
| B.6      | The $F$ distribution . . . . .  | 226        |
| B.7      | The Wishart distribution . . . . .  | 227        |
| B.8      | The Dirichlet distribution . . . . .                                      | 228        |
| <b>C</b> | <b>The language of directed acyclic graphical models</b>                  | <b>230</b> |
| C.1      | Directed graphical models and their interpretation . . . . .              | 230        |
| C.2      | Conditional independence in DAGS: d-separation† . . . . .                 | 233        |
| <b>D</b> | <b>Dummy chapter</b>  | <b>235</b> |

# Chapter 2

## Univariate Probability

This chapter briefly introduces the fundamentals of univariate probability theory, density estimation, and evaluation of estimated probability densities.

### 2.1 What are probabilities, and what do they have to do with language?

We'll begin by addressing a question which is both philosophical and practical, and may be on the minds of many readers: *What are probabilities, and what do they have to do with language?* We'll start with the classic but non-linguistic example of coin-flipping, and then look at an analogous example from the study of language.

#### Coin flipping

You and your friend meet at the park for a game of tennis. In order to determine who will serve first, you jointly decide to flip a coin. Your friend produces a quarter and tells you that it is a fair coin. What exactly does your friend mean by this?

A translation of your friend's statement into the language of probability theory would be that the tossing of the coin is an EXPERIMENT—a repeatable procedure whose outcome may be uncertain—in which the probability of the coin landing with heads face up is equal to the probability of it landing with tails face up, at  $\frac{1}{2}$ . In mathematical notation we would express this translation as  $P(\text{Heads}) = P(\text{Tails}) = \frac{1}{2}$ . This mathematical translation is a partial answer to the question of what probabilities are. The translation is not, however, a complete answer to the question of what your friend means, until we give a semantics to statements of probability theory that allows them to be interpreted as pertaining to facts about the world. This is the philosophical problem posed by probability theory.

Two major classes of answer have been given to this philosophical problem, corresponding to two major schools of thought in the application of probability theory to real problems in the world. One school of thought, the *frequentist* school, considers the probability of an event to denote its limiting, or asymptotic, frequency over an arbitrarily large number of repeated

trials. For a frequentist, to say that  $P(\text{Heads}) = \frac{1}{2}$  means that if you were to toss the coin many, many times, the proportion of Heads outcomes would be guaranteed to eventually approach 50%.

The second, *Bayesian* school of thought considers the probability of an event  $E$  to be a principled measure of the strength of one's belief that  $E$  will result. For a Bayesian, to say that  $P(\text{Heads})$  for a fair coin is 0.5 (and thus equal to  $P(\text{Tails})$ ) is to say that you believe that Heads and Tails are equally likely outcomes if you flip the coin. A popular and slightly more precise variant of Bayesian philosophy frames the interpretation of probabilities in terms of rational betting behavior, defining the probability  $\pi$  that someone ascribes to an event as the maximum amount of money they would be willing to pay for a bet that pays one unit of money. For a fair coin, a rational better would be willing to pay no more than fifty cents for a bet that pays \$1 if the coin comes out heads.<sup>1</sup>

The debate between these interpretations of probability rages, and we're not going to try and resolve it here, but it is useful to know about it, in particular because the frequentist and Bayesian schools of thought have developed approaches to inference that reflect these philosophical foundations and, in some cases, are considerably different in approach. Fortunately, for the cases in which it makes sense to talk about both reasonable belief and asymptotic frequency, it's been proven that the two schools of thought lead to the same rules of probability. If you're further interested in this, I encourage you to read Cox (1946), a beautiful, short paper.

### **An example of probabilities in language: word ordering**

There were two parts to formalizing the notion of probability in the coin-flipping example: (1) delimiting the world of possible outcomes, and (2) assigning probabilities to each possible outcome. Each of these steps involves a simplification. Step 1 ignores such details as the angle between the "vertical" axis of the coin's face and magnetic north which results from the flip, and omits such possibilities as that the coin will land on its edge, that it will be snatched up by an owl, and so forth. Step 2 omits contingent information such as the relative orientation of the coin upon its being flipped, how hard it is flipped, the air currents, and so forth. With these simplifications, however, comes a great deal of analytical traction and power. Cases such as these, in which we can delimit a world of possible outcomes and express probabilities over those outcomes on the basis of incomplete knowledge, are ubiquitous in science, and are also ubiquitous in language. As a simple example analogous to coin flipping, let us consider the choice of how to order the words in an English BINOMIAL (Malkiel, 1959; Cooper and Ross, 1975; Benor and Levy, 2006, *inter alia*), such as *principal and interest*, where both orders are observed in naturally occurring usage. For a linguist to claim that this binomial has no ordering preference can be translated into the language of probability theory as stating that we are equally likely to observe (in some set of contexts of English

---

<sup>1</sup>This definition in turn raises the question of what "rational betting behavior" is. The standard response to this question defines rational betting as betting behavior that will never enter into a combination of bets that is guaranteed to lose money, and will never fail to enter into a combination of bets that is guaranteed to make money. The arguments involved are called "Dutch Book arguments" (Jeffrey, 2004).



usage) the phrases *principal and interest* and *interest and principal*; if we abbreviate these two orderings as *p* and *i* (we might denote the union of the two orderings as  $\{\textit{interest, principal}\}$ ) then mathematically our linguist is saying that  $P(\mathbf{p}) = P(\mathbf{i}) = \frac{1}{2}$ .

## 2.2 Sample Spaces

The underlying foundation of any probability distribution is the SAMPLE SPACE—a set of possible OUTCOMES, conventionally denoted  $\Omega$ . For example, the sample space for orderings of the unordered binomial pair  $\{\textit{principal, interest}\}$  is

$$\Omega = \{\mathbf{p}, \mathbf{i}\} \tag{2.1}$$

If we were to observe two tokens of the binomial, then the sample space would be

$$\Omega = \{\mathbf{pp}, \mathbf{pi}, \mathbf{ip}, \mathbf{ii}\} \tag{2.2}$$

In general, sample spaces can be finite (e.g., the set of all syntactic categories), countably infinite (e.g., the set of integers, the set of all phrase-structure trees), or uncountably infinite (e.g., the set of real numbers).

## 2.3 Events and probability spaces

An EVENT is simply a subset of a sample space. In the interpretation of probability distributions as beliefs, events are often interpreted as PROPOSITIONS.

What is the sample space corresponding to the roll of a single six-sided die? What is the event that the die roll comes up even?

It follows that the negation of an event  $E$  (that is,  $E$  not happening) is simply  $\Omega - E$ .

A PROBABILITY SPACE  $P$  on  $\Omega$  is a function from events in  $\Omega$  to real numbers such that the following three axioms hold:

1.  $P(E) \geq 0$  for all  $E \subset \Omega$  (NON-NEGATIVITY).
2. If  $E_1$  and  $E_2$  are disjoint, then  $P(E_1 \cup E_2) = P(E_1) + P(E_2)$  (DISJOINT UNION).
3.  $P(\Omega) = 1$  (PROPERNESS).

These axioms allow us to express the probabilities of some events in terms of others.

## 2.4 Conditional Probability, Bayes' rule, and Independence

The **CONDITIONAL PROBABILITY** of event  $B$  given that  $A$  has occurred/is known is defined as follows:

$$P(B|A) \equiv \frac{P(A \cap B)}{P(A)}$$

We'll use another type of word-ordering example to illustrate this concept. In Old English, the object in a transitive sentence could appear either preverbally or postverbally. It is also well-documented in many languages that the “weight” of a noun phrase (as measured, for example, by number of words or syllables) can affect its preferred position in a clause, and that pronouns are “light” (Hawkins, 1994; Wasow, 2002). Suppose that among transitive sentences in a corpus of historical English, the frequency distribution of object position and pronominality is as follows:

|                             | <b>Pronoun</b> | <b>Not Pronoun</b> |
|-----------------------------|----------------|--------------------|
| (1) Object <b>Preverbal</b> | 0.224          | 0.655              |
| Object <b>Postverbal</b>    | 0.014          | 0.107              |

For the moment, we will interpret these frequencies directly as probabilities. (We'll see more on this in Chapter 4.) What is the conditional probability of pronominality given that an object is postverbal?

In our case, event  $A$  is **Postverbal**, and  $B$  is **Pronoun**. The quantity  $P(A \cap B)$  is already listed explicitly in the lower-right cell of table I: 0.014. We now need the quantity  $P(A)$ . For this we need to calculate the **MARGINAL TOTAL** of row 2 of Table I:  $0.014 + 0.107 = 0.121$ . We can then calculate:

$$\begin{aligned} P(\mathbf{Pronoun}|\mathbf{Postverbal}) &= \frac{P(\mathbf{Postverbal} \cap \mathbf{Pronoun})}{P(\mathbf{Postverbal})} \\ &= \frac{0.014}{0.014 + 0.107} = 0.116 \end{aligned}$$

### The chain rule

If we have events  $E_1, E_2, \dots, E_n$ , then we can recursively apply the definition of conditional independence to the probability of all these events occurring— $P(E_1 \cap E_2 \cap \dots \cap E_n)$ —to obtain

$$P(E_1 \cap E_2 \cap \dots \cap E_n) = P(E_n|E_1 \cap E_2 \cap \dots \cap E_{n-1}) \dots P(E_2|E_1)P(E_1) \quad (2.3)$$

Equation 2.3 is known as the **CHAIN RULE**, and using it to decompose a complex probability distribution is known as **CHAIN RULE DECOMPOSITION**.

### 2.4.1 Bayes' rule

BAYES' RULE (also called Bayes' theorem) is simply the expression of a conditional probability in terms of the converse conditional probability and the two relevant unconditional probabilities:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (2.4)$$

Bayes' rule can also be extended to more complex conjunctions of events/propositions:

$$P(A|B, I) = \frac{P(B|A, I)P(A|I)}{P(B|I)} \quad (2.5)$$

Although Bayes' rule is a simple mathematical truth, it acquires profound conceptual and practical power when viewed as a way of updating beliefs (encoded as probability distributions) in the face of new information. Specifically, suppose belief in  $A$  is of interest. One's initial, or PRIOR, beliefs in  $A$  are quantified by  $P(A|I)$ . Bayes' rule then expresses how beliefs should change when  $B$  is learned. In particular, the POSTERIOR belief  $P(A|B, I)$  in  $A$  equals the prior belief times the ratio between (i) the LIKELIHOOD  $P(B|A, I)$  of  $B$  under  $A$  and  $I$  and (ii) the likelihood of  $B$  under  $I$  alone. This use of Bayes' rule is often called BAYESIAN INFERENCE, and it serves as the cornerstone of (fittingly) Bayesian statistics.

We will see many examples of Bayesian inference throughout this book, but let us work through a simple example to illustrate its basic workings. We will return to the domain of Old English word order, but now focus on the relationship between an object NP's word order and its ANIMACY (assuming every object is either animate or inanimate) rather than its pronominality. Suppose we have the following probabilities:

$$\begin{aligned} P(\text{Object } \mathbf{Animate}) &= 0.4 & (2.6) \\ P(\text{Object } \mathbf{Postverbal} | \text{Object } \mathbf{Animate}) &= 0.7 \\ P(\text{Object } \mathbf{Postverbal} | \text{Object } \mathbf{Inanimate}) &= 0.8 \end{aligned}$$

and that we want to compute how likely an object is to be animate given that it is expressed postverbally—that is,  $P(\text{Object } \mathbf{Animate} | \text{Object } \mathbf{Postverbal})$  (e.g., a comprehender may know at some point in a sentence that the object will appear postverbally, but hasn't yet heard the object spoken). Although we aren't given a probability table as in Example I, we actually have all the information necessary to compute this probability using Bayes' rule. We go through the calculations step by step below, simplifying the notation by using **Anim** and **Inanim** respectively to denote animacy and inanimacy of the object, and **PreV** and **PostV** respectively to denote preverbal and postverbal positioning of the object.

$$P(\mathbf{Anim} | \mathbf{PostV}) = \frac{P(\mathbf{PostV} | \mathbf{Anim})P(\mathbf{Anim})}{P(\mathbf{PostV})} \quad (2.7)$$

We have been given the value of the two terms in the numerator, but let us leave the numerator alone for the moment and focus on the denominator, which we weren't given in the problem specification. At first, it may not be obvious how to compute the denominator. However, we can Axiom 2 of probability theory (disjoint union) to express  $P(\mathbf{PostV})$  as the sum of the probabilities  $P(\mathbf{PostV} \cap \mathbf{Anim})$  and  $P(\mathbf{PostV} \cap \mathbf{Inanim})$ :

$$P(\mathbf{PostV}) = P(\mathbf{PostV} \cap \mathbf{Anim}) + P(\mathbf{PostV} \cap \mathbf{Inanim})$$

Although these probabilities were not specified directly either, we can use the definition of conditional probability to turn them into forms that *were* specified:

$$\begin{aligned} P(\mathbf{PostV} \cap \mathbf{Anim}) &= P(\mathbf{PostV}|\mathbf{Anim})P(\mathbf{Anim}) \\ P(\mathbf{PostV} \cap \mathbf{Inanim}) &= P(\mathbf{PostV}|\mathbf{Inanim})P(\mathbf{Inanim}) \end{aligned} \tag{2.8}$$

Now we can plug this result back into Equation (2.7):

$$P(\mathbf{Anim}|\mathbf{PostV}) = \frac{P(\mathbf{PostV}|\mathbf{Anim})P(\mathbf{Anim})}{P(\mathbf{PostV}|\mathbf{Anim})P(\mathbf{Anim}) + P(\mathbf{PostV}|\mathbf{Inanim})P(\mathbf{Inanim})} \tag{2.9}$$

At this point, it is worth reflecting on the expanded form Bayes' rule for this problem that we see in Equation (2.9). First, note that we have rewritten the initial form of Bayes' rule into a formula all of whose terms we have immediate access to in the probability specifications given in (2.6). (We were not given  $P(\mathbf{Inanim})$ , but the axioms of probability theory—disjoint union together with properness—allow us to easily determine that its value is 0.6.) Hence we can immediately calculate the correct answer to our problem:

$$P(\mathbf{Anim}|\mathbf{PostV}) = \frac{0.7 \times 0.4}{0.7 \times 0.4 + 0.8 \times 0.6} = 0.3684 \tag{2.10}$$

Second, note that in the right-hand side of Equation (2.9), the numerator appears as one of the two terms being summed in the denominator. This is quite often the case in applications of Bayes' rule. It was the fact that **Anim** and **Inanim** constitute an exhaustive partition of our sample space that allowed us to break down  $P(\mathbf{PostV})$  in the way we did in Equation (2.8). More generally, it is quite common for the most complex step of applying Bayes' rule to be breaking the sample space into an exhaustive partition  $A_1, A_2, \dots, A_n$ , and re-expressing Equation (2.11) through a summation over the members of this exhaustive partition:

$$P(A|B) = \frac{P(B|A)P(A)}{\sum_{j=1}^n P(B|A_j)P(A_j)} \tag{2.11}$$

A closely related third point is that computation of the denominator is usually the most complex and difficult part of applying Bayes' rule. Fortunately, there are often tricks that

one can apply either to avoid this computation or to drastically simplify it; you will see several examples of these tricks later in the book.

Finally, it is worthwhile to compare the probabilities of the object being animate before (Equation (2.6)) versus after (Equation (2.10)) obtaining the knowledge that the object follows the verb. Inspection of these two probabilities reveals that the object is postverbal *reduces* the probability by a small amount, from 0.4 to about 0.37. Inspection of the two conditional probabilities in the problem specification also reveals that inanimate objects are some more likely to be realized postverbally (probability 0.7) than animate objects are (probability 0.8). In fact, the shift induced in probability of object animacy from learning that the object is postverbal directly follows from the differential preference for postverbal realization of inanimate versus animate objects. If the object were inanimate, it would *predict more strongly* than if the object were animate that the object should be postverbal. Hence, learning that the object is in fact postverbal goes some way toward disconfirming the possibility that the object may turn out to be animate, while strengthening the possibility that it may turn out to be inanimate.

### 2.4.2 (Conditional) Independence

Events  $A$  and  $B$  are said to be **CONDITIONALLY INDEPENDENT GIVEN INFORMATION  $C$**  if

$$P(A \cap B|C) = P(A|C)P(B|C) \quad (2.12)$$

This form of conditional independence is often denoted symbolically as  $A \perp B \mid C$ .

A more philosophical way of interpreting conditional independence is that if we are in the state of knowledge denoted by  $C$ , then conditional independence of  $A$  and  $B$  means that knowing  $A$  tells us nothing more about the probability of  $B$ , and vice versa. The simple statement that  $A$  and  $B$  are **CONDITIONALLY INDEPENDENT** is often used; this should be interpreted that  $A$  and  $B$  are conditionally independent given an implicit state  $C$  of “not knowing anything at all” ( $C = \emptyset$ ).

It’s crucial to keep in mind that if  $A$  and  $B$  are conditionally dependent given  $C$ , that does not guarantee they will not be conditionally independent given some other set of knowledge  $D$ . As an example, suppose that your friend gives you a pouch with three coins of identical shape. One coin is two-headed, one coin is two-tailed, and one coin is a regular fair coin; this information constitutes your state of knowledge  $C$ . You are to randomly select a coin from the pouch and, without inspecting it, flip it twice; the outcomes of the flips correspond to events  $A$  and  $B$ . Given this state of affairs,  $A$  and  $B$  are clearly *not* conditionally independent given  $C$ . For example,  $P(B|A = \text{Heads}, C) > P(B|C)$ : knowing that  $A = \text{Heads}$  rules out the third coin and therefore makes it more likely that the second coin flip  $B$  will also come out heads. Suppose, however, that you inspect the coin before flipping it twice; call the new state of knowledge obtained after inspecting the coin  $D$ . We *do* have  $A \perp B \mid D$ : the conditional dependence between  $A$  and  $B$  given  $C$  derived from the uncertainty as to which of the three coins you selected, and once that uncertainty is removed, the dependency is broken and independence is obtained.

Likewise, it is also possible (though less common in real-world circumstances) for conditional independence between events to be *lost* when new knowledge is gained—see Exercise 2.2.

## 2.5 Discrete random variables and probability mass functions

A DISCRETE RANDOM VARIABLE  $X$  is literally a function from the sample space  $\Omega$  of a probability space to a finite, or countably infinite, set of real numbers ( $\mathbb{R}$ ).<sup>2</sup> Together with the function  $P$  mapping elements  $\omega \in \Omega$  to probabilities, a random variable determines a PROBABILITY MASS FUNCTION  $P(X(\omega))$ , or  $P(X)$  for short, which maps real numbers to probabilities. For any value  $x$  in the range of the random variable  $X$ , suppose that  $A$  is the part of the sample space all of whose members  $X$  maps to  $x$ . The probability that  $X$  will take on the value  $x$  is therefore simply the value that the original probability function assigns to  $A$ :

$$P(X = x) = P(A)$$

Technically speaking, the two  $P$ 's in this equation are different—the first applies to values in the range of the random variable  $X$ , whereas the second applies to subsets of the sample space. But

The relationship between the sample space  $\Omega$ , a probability space  $P$  on  $\Omega$ , and a discrete random variable  $X$  on  $\Omega$  can be a bit subtle, so we'll illustrate it by returning to our example of collecting two tokens of *{principal, interest}*. Once again, the sample space is  $\Omega = \{\text{pp, pi, ip, ii}\}$ . Consider the function  $X$  that maps every possible pair of observations—that is, every point in the sample space—to *the total number of p outcomes obtained*. Suppose further that there is no ordering preference for the binomial, so that for each point  $\omega$  in the sample space we have  $P(\{\omega\}) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$ . The total number of **p** outcomes is a random variable  $X$ , and we can make a table of the relationship between  $\omega \in \Omega$ ,  $X(\omega)$ , and  $P(X)$ :

| $\omega$ | $X(\omega)$ | $P(X)$        |
|----------|-------------|---------------|
| pp       | 2           | $\frac{1}{4}$ |
| pi       | 1           | $\frac{1}{2}$ |
| ip       | 1           | $\frac{1}{2}$ |
| ii       | 0           | $\frac{1}{4}$ |

Notice that the random variable  $X$  serves to *partition* the sample space into equivalence classes: for each possible real number  $y$  mapped to by  $X$ , all elements of  $\Omega$  mapped to  $y$  are in that equivalence class. Intuitively, a random variable can be thought of as focusing

---

<sup>2</sup>A set  $S$  is COUNTABLY INFINITE if a one-to-one mapping exists between the integers  $0, 1, 2, \dots$  and  $S$ .

attention on only the distinctions within the sample space that are of interest to us for a particular application. In the example above, the sample space consisted of ordered pairs of  $\{\textit{principal, interest}\}$  binomials, but the random variable  $X$  restricts our attention to the observed *frequency* of each of the two binomial forms, throwing out the information about which binomial is observed first and which is observed second.

### 2.5.1 Bernoulli trials and the Bernoulli distribution

Perhaps the simplest interesting kind of event space is one that contains two outcomes, which we will arbitrarily label “success” and “failure” and respectively associate the integers 1 and 0. A **BERNOULLI TRIAL** is an experiment (in the sense of Section 2.1) with these two possible outcomes. This leads us to our first **PARAMETRIC FAMILY OF PROBABILITY DISTRIBUTIONS**, the **BERNOULLI DISTRIBUTION**. A parametric family of probability distributions is an infinite collection of probability distributions that vary only in the value of a fixed number of **PARAMETERS** characterizing the family. The Bernoulli distribution is perhaps the simplest of these families, being characterized by a single parameter, which we will denote by  $\pi$ .  $\pi$  is the probability of achieving success on a single Bernoulli trial, and can take any value between 0 and 1 (inclusive);  $\pi$  is sometimes called the “success parameter”. The Bernoulli distribution thus has a probability mass function of the form

$$P(X = x) = \begin{cases} \pi & \text{if } x = 1 \\ 1 - \pi & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases}$$

A random variable that follows a Bernoulli distribution is often called a **BERNOULLI RANDOM VARIABLE**. For example, the flipping of a fair coin (with heads mapped to 1 and tails to 0) or the ordering outcome of an English binomial with no ordering preference can be modeled as a Bernoulli random variable with parameter  $\pi = 0.5$ .

### 2.5.2 Multinomial trials

We can also generalize the Bernoulli trial to the case where there are  $r \geq 2$  possible outcomes; for convenience we can label the outcomes  $c_1, \dots, c_r$ . This is a **MULTINOMIAL TRIAL**, and just as the Bernoulli distribution has 1 parameter, the distribution for multinomial trials has  $r - 1$  parameters  $\pi_1, \dots, \pi_{r-1}$  determining the probability that a trial will fall into each of the classes:

$$P(X = x) = \begin{cases} \pi_1 & \text{if } x = c_1 \\ \pi_2 & \text{if } x = c_2 \\ \vdots & \vdots \\ \pi_{r-1} & \text{if } x = c_{r-1} \\ 1 - \sum_{i=1}^{r-1} \pi_i & \text{if } x = c_r \\ 0 & \text{otherwise} \end{cases} \quad (2.13)$$

We can make Equation (2.13) more symmetric by defining  $\pi_r \stackrel{\text{def}}{=} 1 - \sum_{i=1}^{r-1} \pi_i$ , which allows us to replace the second-to-last line of 2.13 with

$$P(X = x) = \pi_r \quad \text{if } x = c_r$$

but you should keep in mind that  $\pi_r$  is not an independent parameter, since it is fully determined by the other parameters.

**Example.** You decide to pull *Alice in Wonderland* off your bookshelf, open to a random page, put your finger down randomly on that page, and record the letter that your finger is resting on (ignoring the outcome and trying again if your finger rests on punctuation or a space). This procedure can be modeled as a multinomial trial with 26 possible outcomes, and to a first approximation the parameters of the associated distribution can simply be the relative frequencies of the different letters (ignoring the differing widths and heights of the letters). In *Alice in Wonderland*, 12.6% of the letters are **e**, 9.9% are **t**, 8.2% are **a**, and so forth; so we could write the parameters of our model as  $\pi_e = 0.126$ ,  $\pi_t = 0.099$ ,  $\pi_a = 0.082$ , and so forth.

The probability distribution for multinomial trials discussed here is a special case of the more general MULTINOMIAL DISTRIBUTION introduced in Section 3.4.1.

## 2.6 Cumulative distribution functions

A random variable  $X$  determines a probability mass function  $P(X)$  on the real numbers. This probability mass function in turn determines a CUMULATIVE DISTRIBUTION FUNCTION  $F$ , defined as

$$F(x) \stackrel{\text{def}}{=} P(X \leq x)$$

We give a very simple illustration with the Bernoulli distribution. A Bernoulli random variable with parameter  $\pi$  has the following (very simple!) cumulative distribution function:



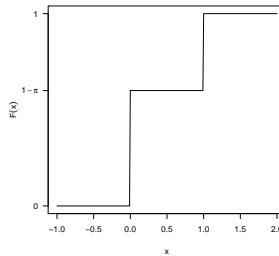


Figure 2.1: The cumulative distribution function for a Bernoulli random variable with parameter  $\pi$

$$F(x) = \begin{cases} 0 & x < 0 \\ \pi & 0 \leq x < 1 \\ 1 & x \geq 1 \end{cases}$$

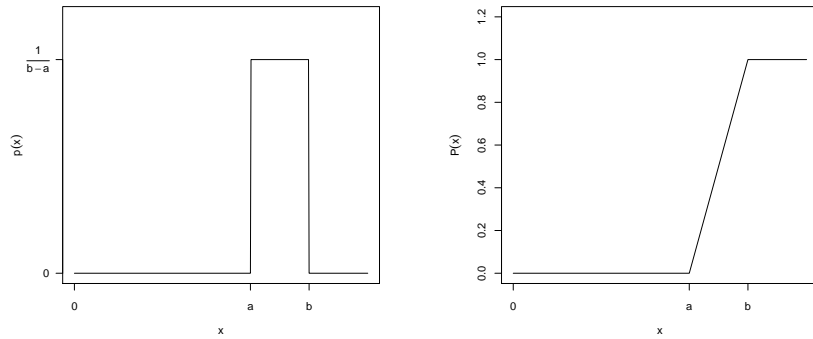
which is illustrated in Figure 2.1.

A probability mass function uniquely determines a cumulative distribution function, and vice versa.

Note that the cumulative distribution function is monotonically increasing in the range of the random variable. This means that the cumulative distribution function has an *inverse*, the QUANTILE FUNCTION, which maps a probability  $0 \leq p \leq 1$  into the lowest possible number  $x$  such that  $P(X \leq x) \geq p$ .

## 2.7 Continuous random variables and probability density functions

Limiting a random variable to take on at most a countably infinite set of values is often too strong a constraint: in many cases, we want to allow outcomes to range along a continuum of real values, which is uncountably infinite. This type of outcome requires different treatment, with CONTINUOUS RANDOM VARIABLES. Instead of a discrete random variable's probability mass function, a continuous random variable has a PROBABILITY DENSITY FUNCTION  $p(x)$  that assigns non-negative density to every real number. For example, the amount of time that an infant lives before it hears a parasitic gap in its native language would be naturally modeled as a continuous random variable (with  $p(x) > 0$  only for  $x > 0$ ). If we plot the probability density function (pdf) as a curve over the real number line, then the properness requirement of probability theory ensures that the total area under the curve is equal to 1 (see example in Section 2.7.1):



(a) Probability density function      (b) Cumulative distribution function

Figure 2.2: The probability density function and cumulative distribution function of the uniform distribution with parameters  $a$  and  $b$

$$\int_{-\infty}^{\infty} p(x) dx = 1$$

### 2.7.1 The uniform distribution

The simplest parametrized family of continuous probability distributions is the **UNIFORM DISTRIBUTION**, defined by parameters  $a$  and  $b$  bounding a continuous region  $[a, b]$  within which the density function  $p(x)$  is constant, and outside of which  $p(x) = 0$ . Since the area under the pdf curve must total 1, we must have the probability density function

$$P(x|a, b) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases} \quad (2.14)$$

$p(x) = \frac{1}{b-a}$  when  $x \in [a, b]$ . We sometimes denote that a random variable  $X$  is distributed uniformly on the interval  $[a, b]$  with the following notation:

$$X \sim \mathcal{U}(a, b)$$

Figure 2.2 shows plots of the pdf and cdf of the uniform distribution.

**Example:** although the uniform distribution is the simplest example of a continuous probability distribution, it is not the continuous distribution that has the most obvious or widespread applications in linguistics. One area in which uniform distributions would be most applicable, however, is historical applications, particularly as pertains to inferring event times such as the dates of recovered documents or of divergence between related languages. The written language of Ancient Aramaic, for example, is known to have been in use from roughly

1000 B.C.E. to 500 B.C.E. (Beyer, 1986). With only this information, a crude estimate of the distribution over Ancient Aramaic document dates might be a uniform distribution on the interval  $[-1000, -500]$  (though this estimate would fail to incorporate additional information likely available from the documents themselves). Suppose a scholar consults two documents of Ancient Aramaic from unrelated sources, and finds that the date of the first document is 850 B.C.E. What is the probability that the other document dates to within fifty years of the first? Anything in the range  $[-900, -800]$  qualifies, so the probability we want is:

$$\begin{aligned} P(X \in [-900, -800]) &= \int_{-900}^{-800} \frac{1}{-500 - (-1000)} dx \\ &= \frac{-800 - (-900)}{-500 - (-1000)} = \frac{1}{5} \end{aligned}$$

## 2.7.2 Cumulative distribution functions for continuous random variables

With a continuous random variable, the probability of any specific point in  $\mathbb{R}$  is zero; the primary interest is on the probability that the outcome of the random variable will fall into a given *region* of  $\mathbb{R}$ , which is more naturally expressed via the cumulative distribution function (cdf)  $F(x)$ , defined once again as  $P(X \leq x)$ , or

$$F(x) = \int_{-\infty}^x p(x) dx$$

What is especially useful about the cumulative distribution function is that the probability that  $X$  will fall into a continuous *region*  $[x, y]$  of the real number line can be expressed as the difference between the cdf at  $y$  and at  $x$ :

$$\begin{aligned} P(x \leq X \leq y) &= \int_x^y p(x) dx \\ &= \int_{-\infty}^y p(x) dx - \int_{-\infty}^x p(x) dx \\ &= F(y) - F(x) \end{aligned}$$

Because of this property, and the fact that the probability of an outcome occurring at any specific point  $x$  is 0 for a continuous random variable, the cumulative distribution is in many cases more important than the density when working with continuous random variables.

## 2.8 Normalized and unnormalized probability distributions

It is quite common to wind up defining a “probability” mass function  $F$  (or density function  $f$ ) that adheres to the first two axioms listed in Section 2.3—non-negativity and disjoint union—but that does not adhere to the third axiom of properness. Such a function  $F$  is called an UNNORMALIZED or IMPROPER PROBABILITY DISTRIBUTION. In these cases, from  $F$  a normalized, or proper, probability distribution  $P$  can be defined as

$$P(X = x) = \frac{1}{Z}F(x) \tag{2.15}$$

where

$$Z \stackrel{\text{def}}{=} \sum_x F(x)$$

for discrete densities, and

$$Z \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} f(x) dx$$

for continuous densities, when  $Z$  is finite. Here,  $Z$  is generally called the NORMALIZING CONSTANT or PARTITION FUNCTION.

**Example:** suppose that we wanted to construct a probability distribution over total orderings of three constituents  $S, O, V$  (e.g., the subject, object, and verb of a simple transitive clause), and characterize the probability distribution purely on the basis of the relative strengths of preference for linear precedence between every possible pair of constituents. There are three possible pairs of these constituents— $SO, SV$ , and  $OV$ —so we will introduce one parameter for each of these pairs to indicate the relative strength of preference for one linear ordering versus another. Thus we will have one parameter,  $\gamma_1$  indicating the preference for  $S$  to precede  $O$ ; we will denote this event as  $S \prec O$ , with  $\prec$  to be read as “precedes”. A second parameter  $\gamma_2$  will indicate the preference for  $S$  to precede  $V$ , and a third parameter  $\gamma_3$  indicating the preference for  $O$  to precede  $V$ . For simplicity, we’ll let each  $\gamma_i$  range between 0 and 1, and encourage the intuitive analogy between these three parameters and the success parameters of three separate Bernoulli trials. That is, the word orders  $SOV, SVO$ , and  $VSO$  could be thought of as “successes” for the Bernoulli trial, and we would want them together to have something like probability  $\gamma_1$ , whereas the word orders  $OSV, OVS$ , and  $VOS$  could be thought of as “failures” for this Bernoulli trial, and we would want them to have something like probability  $1 - \gamma_1$ . So we’ll define a mass function  $F$  that assigns to any word order the product of (i) all the “success” parameters for precedences that are satisfied, and (ii) all the “failure” parameters for precedences that are violated. For example, we would have:

$$F(SOV) = \gamma_1\gamma_2\gamma_3$$

since this word ordering satisfies all the three precedences,

$$F(SVO) = \gamma_1\gamma_2(1 - \gamma_3)$$

since this word ordering violates  $O \prec V$  but satisfies the other two precedences, and so forth.

However, there is one crucial disanalogy between the present case and the case of three separate Bernoulli trials. In three separate Bernoulli trials, there are eight logically possible outcomes, but in the case of ordering three constituents, there are only six logically possible outcomes. There are two combinations of constituent-pair precedence which are contradictory:

$$(1) S \prec O, V \prec S, O \prec V$$

$$(2) O \prec S, S \prec V, V \prec O$$

As a result, the mass function  $F$  is improper: in general, it does not assign total mass of 1 to the six possible constituent orderings.

We can construct a proper probability distribution out of  $F$ , however, by computing its normalizing constant and then defining a probability distribution as described in Equation 2.15. Below we make totally explicit the eight logically possible combinations of three pairwise precedences, the two that are contradictory, and the values assigned by  $F$  to each of the eight.

|   | $S\_O$  | $S\_V$  | $O\_V$  | Outcome $X$       | $F(X)$                                       |
|---|---------|---------|---------|-------------------|--|
| 1 | $\prec$ | $\prec$ | $\prec$ | $SOV$             | $\gamma_1\gamma_2\gamma_3$                   |
| 2 | $\prec$ | $\prec$ | $\succ$ | $SVO$             | $\gamma_1\gamma_2(1 - \gamma_3)$             |
| 3 | $\prec$ | $\succ$ | $\prec$ | <b>impossible</b> | $\gamma_1(1 - \gamma_2)\gamma_3$             |
| 4 | $\prec$ | $\succ$ | $\succ$ | $VSO$             | $\gamma_1(1 - \gamma_2)(1 - \gamma_3)$       |
| 5 | $\succ$ | $\prec$ | $\prec$ | $OSV$             | $(1 - \gamma_1)\gamma_2\gamma_3$             |
| 6 | $\succ$ | $\prec$ | $\succ$ | <b>impossible</b> | $(1 - \gamma_1)\gamma_2(1 - \gamma_3)$       |
| 7 | $\succ$ | $\succ$ | $\prec$ | $OVS$             | $(1 - \gamma_1)(1 - \gamma_2)\gamma_3$       |
| 8 | $\succ$ | $\succ$ | $\succ$ | $VOS$             | $(1 - \gamma_1)(1 - \gamma_2)(1 - \gamma_3)$ |

Since the set of all eight together would give a proper distribution, a simple way of expressing the normalizing constant is as  $1 - F(X_3) - F(X_6)$ .

Problem 2.6 in the end of this chapter revisits the question of whether the parameters  $\gamma_i$  really turn out to be the probabilities of satisfaction or violation of each individual constituent-pair precedence relation for the probability distribution resulting from our choice of the unnormalized mass function  $F$ .

When we have a function  $F(X)$  that we wish to use as an unnormalized probability distribution, we will often write that

$$P(x) \propto F(x) \tag{2.16}$$

which is read as “ $P(x)$  is proportional to  $F(x)$ ”.

## 2.9 Expected values and variance

We now turn to two fundamental quantities of probability distributions: EXPECTED VALUE and VARIANCE.

### 2.9.1 Expected value

The expected value of a random variable  $X$ , which is denoted in many forms including  $E(X)$ ,  $E[X]$ ,  $\langle X \rangle$ , and  $\mu$ , is also known as the EXPECTATION or MEAN. For a discrete random variable  $X$  under probability distribution  $P$ , it’s defined as

$$E(X) = \sum_i x_i P(X = x_i) \tag{2.17}$$

For a Bernoulli random variable  $X$  with parameter  $\pi$ , for example, the possible outcomes are 0 and 1, so we have

$$\begin{aligned} E(X) &= 0 \times \overbrace{(1 - \pi)}^{P(X=0)} + 1 \times \overbrace{\pi}^{P(X=1)} \\ &= \pi \end{aligned}$$

For a continuous random variable  $X$  under cpd  $p$ , the expectation is defined using integrals instead of sums, as

$$E(X) = \int_{-\infty}^{\infty} x p(x) dx \tag{2.18}$$

For example, a uniformly-distributed random variable  $X$  with parameters  $a$  and  $b$  has expectation right in the middle of the interval, at  $\frac{a+b}{2}$  (see Exercise 2.7).

### 2.9.2 Variance

The variance is a measure of how broadly distributed the r.v. tends to be. It’s defined as the expectation of the squared deviation from the mean:

$$\text{Var}(X) = E[(X - E(X))^2] \tag{2.19}$$

The variance is often denoted  $\sigma^2$  and its positive square root,  $\sigma$ , is known as the STANDARD DEVIATION.

## Variance of Bernoulli and uniform distributions

The variance of a Bernoulli-distributed random variable needs to be calculated explicitly, by using the definition in Equation (2.19) and summing over the possible outcomes as in Equation (2.17) (recall that the expectation for a Bernoulli random variable is  $\pi$ ):

$$\begin{aligned}\text{Var}(X) &= E[((X) - E(X))^2] = \sum_{x \in \{0,1\}} (x - \pi)^2 P(x) \\ &= (\pi - 0)^2 \overbrace{(1 - \pi)}^{P(X=0)} + (1 - \pi)^2 \times \overbrace{\pi}^{P(X=1)} \\ &= \pi(1 - \pi) [\pi + (1 - \pi)] \\ \text{Var}(X) &= \pi(1 - \pi)\end{aligned}$$

Note that the variance is largest at  $\pi = 0.5$  and zero when  $\pi = 0$  or  $\pi = 1$ .

The uniform distribution also needs its variance explicitly calculated; its variance is  $\frac{(b-a)^2}{12}$  (see Exercise 2.8).

## 2.10 The normal distribution

We're now in a position to introduce the NORMAL DISTRIBUTION, which is likely to be the most common continuous distribution you'll encounter. It is characterized by two parameters, the expected value  $\mu$  and the variance  $\sigma^2$ . (Sometimes the STANDARD DEVIATION  $\sigma$  is used instead of the variance to parameterize the normal distribution.) Its probability density function is:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(x - \mu)^2}{2\sigma^2} \right] \quad (2.20)$$

This expression seems intimidating at first glance, but it will become familiar with time.<sup>3</sup> It can be divided into three components:

- $\frac{1}{\sqrt{2\pi\sigma^2}}$  is a normalizing constant (see Section 2.8).
- The denominator within the exponential,  $2\sigma^2$ , can be thought of a scaling factor determined by the variance of the normal distribution.
- The numerator within the exponential,  $(x - \mu)^2$ , is the square of the Euclidean distance of  $x$  from the mean. The exponent is negative, so the probability density is exponentially decreasing in the square of the distance from the mean.

---

<sup>3</sup> $\exp[x]$  is another way of writing  $e^x$ ; it's used when the expression in the exponent is complex enough to warrant typesetting in normal-size font.

The normal distribution doesn't have a closed-form cumulative density function, but the approximate cumulative density can be calculated numerically and is available in most statistical software packages. Figure 2.3 shows probability density and cumulative distribution functions for normal distributions with different means and variances.

**Example:** normal distributions are often used for modeling the variability in acoustic dimensions for production and perception in phonetics. Suppose that you are about to record an adult male native speaker of American English pronouncing the vowel [i]. Data from Peterson and Barney (1952) indicate that the F1 formant frequency for this vowel as pronounced by this group may reasonably be modeled as normally distributed with mean 267Hz and standard deviation 36.9Hz. What is the probability that the recording will have F1 frequency falling between 225Hz and 267Hz (lower than average but not egregiously low)? We follow the logic of Section 2.7.2 in expressing the answer in terms of the cumulative distribution function:

$$P(225\text{Hz} \leq \text{F1} \leq 267\text{Hz}) = \int_{225}^{267} p(x) \, dx \quad (2.21)$$

$$= \int_{-\infty}^{267} p(x) \, dx - \int_{-\infty}^{225} p(x) \, dx \quad (2.22)$$

$$= F(267) - F(225) \quad (2.23)$$

With the use of standard statistical software we can find the values of the cumulative distributions function at 267 and 225, which gives us our answer:

$$= 0.5 - 0.12 = 0.38 \quad (2.24)$$

(Note that because the normal distribution is symmetric around its mean, the cumulative distribution function applied to the mean will always be equal to 0.5.)

### 2.10.1 Standard normal random variables

A normal distribution with mean 0 and variance 1 is called the STANDARD NORMAL DISTRIBUTION, and a random variable following this distribution is called a STANDARD NORMAL RANDOM VARIABLE. The density function for a standard normal random variable is  $p(x) = \frac{1}{\sqrt{2\pi}} e^{[-x^2/2]}$ .

## 2.11 Estimating probability densities

Thus far we have concerned ourselves with definitions in probability theory and a few important probability distributions. Most of the time, however, we are in the situation of not knowing the precise distribution from which a set of data have arisen, but of having to infer a probability distribution from the observed data. This is the topic of STATISTICAL INFERENCE. We conclude this chapter by briefly describing some simple techniques for estimating probability densities and evaluating the quality of those estimated densities.



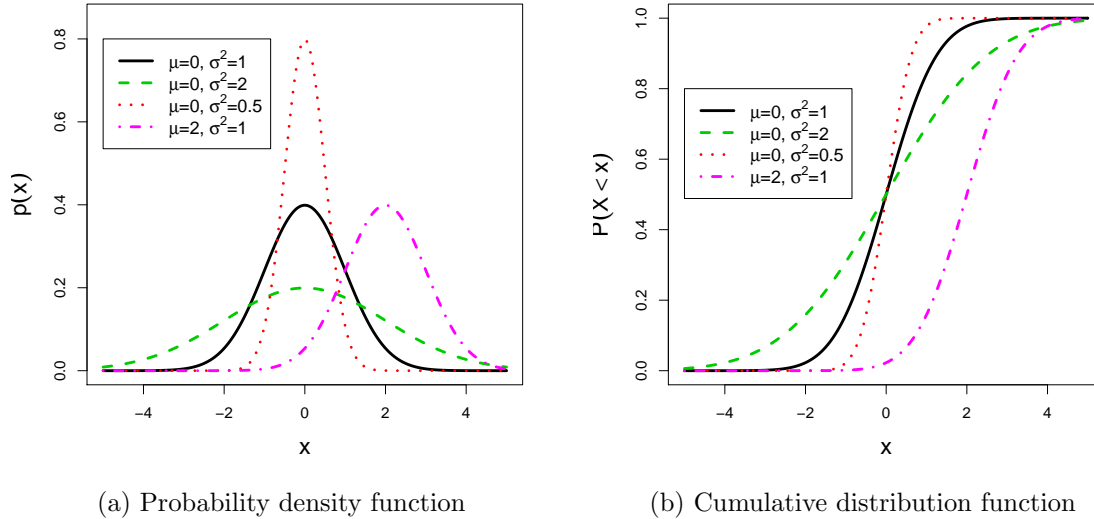


Figure 2.3: The normal distribution: density and cumulative distribution functions

### 2.11.1 Discrete probability densities: relative-frequency estimation

Suppose that we are interested in estimating the Bernoulli parameter  $\pi$  associated with the ordering preference of the English binomial  $\{interest, principal\}$  on the basis of ten tokens collected from a corpus, with “success” arbitrarily associated with the ordering *principal and interest*. Seven of them are *principal and interest* and three are *interest and principal*. The simplest way for us to estimate the underlying Bernoulli distribution is to equate relative frequency of occurrence with probability. In this case, we have seven of ten success so we estimate  $\hat{\pi} = \frac{7}{10}$ .<sup>4</sup> This process is called **RELATIVE FREQUENCY ESTIMATION**.

We can generalize relative frequency estimation for use with multinomial trials. Suppose that we observe  $N$  outcomes of a categorical variable that can take on some finite number  $r$  of different values. Table 2.1, for example, shows the counts of full NP pairs (presented in the order in which they appear in the clause) obtained in a sample of a corpus of German newspaper text.

In terms of probability theory, this categorical variable can be viewed as a discrete multinomial-trial random variable, for which we have observed  $N$  outcomes (570, in the case of Table 2.1). Once again we simply divide the count of each outcome by the total count, as shown in the bottom line of Table 2.1. We will see in Section 4.3.1 that, in addition to being highly intuitive, relative frequency estimation for multinomial outcomes has a deep theoretical justification.

<sup>4</sup>The  $\hat{\pi}$  symbol above the  $\pi$  indicates that this is an *estimate* of the parameter  $\pi$  which may or not be the true underlying parameter value.

|                | $\langle \mathbf{SB}, \mathbf{DO} \rangle$ | $\langle \mathbf{SB}, \mathbf{IO} \rangle$ | $\langle \mathbf{DO}, \mathbf{SB} \rangle$ | $\langle \mathbf{DO}, \mathbf{IO} \rangle$ | $\langle \mathbf{IO}, \mathbf{SB} \rangle$ | $\langle \mathbf{IO}, \mathbf{DO} \rangle$ | Total |
|----------------|--|--|--|--|--|--|-------|
| Count          | 478  | 59   | 1  | 3  | 20   | 9  | 570   |
| Relative Freq. | 0.839                                      | 0.104                                      | 0.001                                      | 0.005                                      | 0.035                                      | 0.016                                      |       |

Table 2.1: Frequency of grammatical functions on ordered pairs of full NPs in German newspaper text, drawn from the NEGRA-II corpus (Kempen and Harbusch, 2004). **SB** denotes “subject”, **DO** denotes “direct object”, and **IO** denotes indirect object.

### 2.11.2 Estimating continuous densities: histograms and kernel density estimation

What about for continuous variables? Figure 2.4a plots the frequency of occurrence of F0 formant frequency of the vowel *a* by adult male speakers in the classic study of Peterson and Barney (1952).<sup>5</sup> It is immediately apparent that relative frequency estimation is not suitable for continuous densities; and if we were to treat the distribution over F0 formants as a discrete distribution, we would run into the problem of data sparsity. For example, we would estimate that  $P(\text{F0} = 119\text{Hz}) = 0$  even though there are many observations close to 119Hz.

One common technique for continuous density estimation is the use of HISTOGRAMS. Constructing a histogram involves dividing the range of the random variable into  $K$  equally-spaced bins and counting the number of observations that fall into each bin;  $K$  can be chosen as seems appropriate to the dataset. These counts can then be normalized by the total number of observations to achieve an estimate of the probability density. Figures 2.4b and 2.4c show histograms of adult male speaker F0 frequency for 38-bin histograms of width 5Hz, starting at 95Hz and 94Hz respectively. Although the histogram determines a valid continuous density, it has two weaknesses. First, it assigns zero probability to a number of intervals for which the data seem to suggest possible outcomes (e.g., the 150–155Hz interval in Figure 2.4b, and the 245–250Hz interval in Figure 2.4c). Second, the shape of the histogram is quite sensitive to the exact positioning of the bins—this is apparent in the substantially different shape of the two histograms in the 100–150Hz range.

A generally preferable approach is KERNEL DENSITY ESTIMATION. A KERNEL is simply a weighting function that serves as a measure of the relevance of each observation to any given point of interest on the range of a random variable. Technically, a kernel  $K$  simply takes an observation  $x_i$  and returns a non-negative function  $K(x_i, \cdot)$  which distributes a total probability mass of 1 over the range of the random variable.<sup>6</sup> Hence we have

$$\sum_x K(x_i, x) = 1 \tag{2.25}$$

<sup>5</sup>The measurements reported in this dataset are rounded off to the nearest Hertz, so in Figure 2.4a they are jittered to break ties.

<sup>6</sup>Although a kernel serves as a type of distance metric, it is not necessarily a true distance; in particular, it need not observe the triangle inequality.

in the discrete case, or

$$\int_x K(x_i, x) dx = 1 \tag{2.26}$$

in the continuous case. If one has only a single observation  $x_1$  of the outcome of a random variable  $X$ , then the kernel density estimate of the probability density over  $X$  is simply  $P(X = x) = K(x_1, x)$ . In general, if one has  $n$  observations  $x_1, \dots, x_n$ , then the kernel density estimate for a point  $x$  is the *average* of the densities assigned to  $x$  by the kernel function obtained from each observation:

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^n K(x_i, x) \tag{2.27}$$

It is up to the researcher to choose the particular kernel function. Here, we give an example for a continuous random variable; in the next section we give an example for discrete kernel density estimation.

For continuous random variables, the `NORMAL KERNEL` is perhaps the most popular kernel; for an observation  $x_i$  it simply allocates its probability mass according to a normal density function with mean  $x_i$  and standard deviation  $\sigma$ . The standard deviation is sometimes called the `BANDWIDTH` of the kernel and denoted as  $b$ . The kernel density estimate from a set of observations  $x_1, \dots, x_n$  using bandwidth  $b$  would be:

$$\hat{p}(X = x) = \frac{1}{n\sqrt{2\pi}b^2} \sum_{i=1}^n \exp \left[ -\frac{(x - x_i)^2}{2b^2} \right]$$

Figure 2.4d shows the kernel density estimate of adult male-speaker F0 frequency distribution for  $b = 5$ , with the observations themselves superimposed on the F0-axis (just like Figure 2.4a). Note that the kernel density estimate gives non-zero probability density to the entire number line, and it is visibly non-zero for the entire span between the lowest and highest observations; yet much of the nuance of the data’s empirical distribution is still retained.

Of course, there are many other possible kernel functions for continuous data than the normal kernel. Another simple example is the `RECTANGULAR` kernel, which for an observation  $x_i$  distributes a probability mass of 1 through a uniform distribution centered on  $x_i$  with width  $b$ . Figure 2.4e shows the result of applying this kernel with bandwidth  $b = 3$  to the same F0 data.<sup>7</sup>

One of the difficulties in kernel density estimation is the choice of bandwidth. This choice is ultimately up to the researcher, but in the next section we will introduce some principles

---

<sup>7</sup>In case you try using the `density()` function in R to do rectangular kernel density estimation, the bandwidth is defined differently—as the standard deviation of the kernel’s density function—and you need to adjust the chosen bandwidth accordingly.

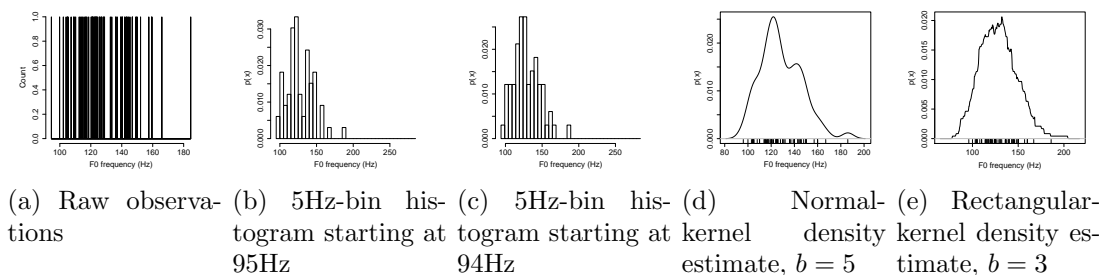


Figure 2.4: Adult male native English speaker F0 measurements for the vowel  $\alpha$  from Peterson and Barney (1952), together with histograms and a kernel density estimate of the underlying distribution

that can help determine how good a given choice of bandwidth may be. Exercise 2.10 also addresses this issue.

Histogram- and kernel-based density estimation is often called **NON-PARAMETRIC ESTIMATION**. The term “non-parametric” turns up in many places in probability and statistics. In density estimation, a non-parametric method is one whose estimated densities can grow arbitrarily complex as the amount of data used for estimation continues to grow. This contrasts with **PARAMETRIC ESTIMATION**, in which estimation is limited to a pre-specified parametric family of models (covered in Chapter 4).

### 2.11.3 Kernel estimation for discrete densities

Kernel density estimation can also be useful for estimating discrete probability distributions in which the number of possible outcomes is large in comparison to the number of observations, or even countably infinite. This is a common situation in the study of language. For example, there has been considerable recent interest in estimating probability densities over possible phonological forms for lexical items (e.g., Hayes and Wilson, 2007), to account for phenomena such as gradience in speaker judgments of nonce word well-formedness. One way of placing a density over the set of possible phonological forms is to define a kernel over pairs of phoneme sequences.

As a simple example, let us consider the space of possible consonant-vowel-consonant (CVC) lexical forms composed of the six phonemes  $/t/$ ,  $/p/$ ,  $/k/$ ,  $/\text{æ}/$ ,  $/\text{ʌ}/$ , and  $/v/$ . There are 27 possible such forms, and 18 of them occur in the English lexicon. Let us base our kernel on the string-edit distance  $D(x, x')$  between forms  $x$  and  $x'$ , which for these three-phoneme sequences is simply the number of positions at which two strings differ—for example,  $D(/p\text{æ}t/, /t\text{v}t/) = 2$ . Less similar phoneme sequences should have a lower score in our kernel, so let us define our kernel as

$$K(x, x') \propto \frac{1}{(1 + D(x, x'))^3}$$

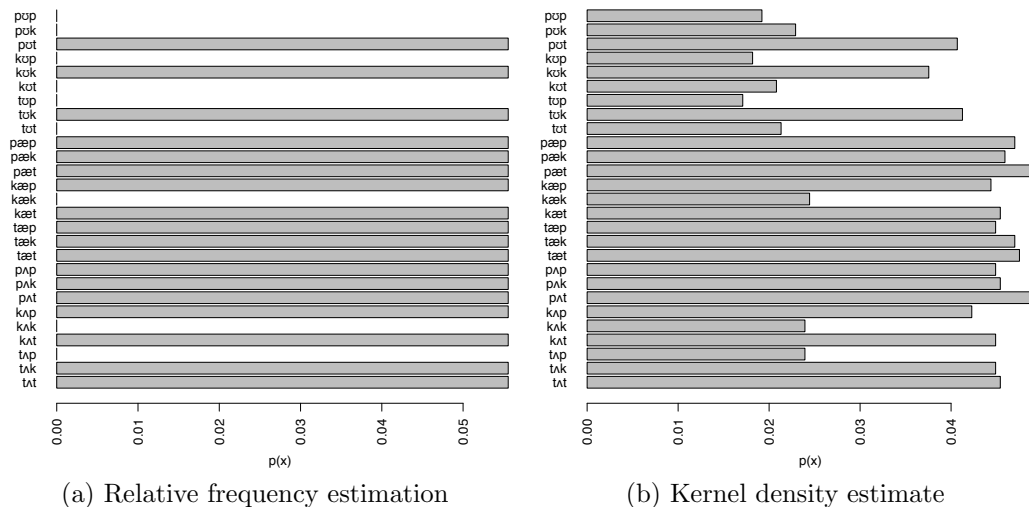


Figure 2.5: Kernel density estimation for lexical forms

We have used proportionality rather than equality here because of the requirement (Equation (2.25)) that the kernel sum to 1 over the complete space of possible outcomes  $\{x_j\}$ . We can renormalize the kernel just as we renormalize a probability distribution (see Section 2.8), by defining for each observation  $x_i$  a normalizing coefficient  $Z_i = \sum_{x_j} \frac{1}{(1+D(x_i, x_j))^3}$ , and dividing in this normalizing coefficient:

$$K(x_i, x') = \frac{1}{Z_i} \frac{1}{(1 + D(x, x'))^3}$$

For our example, it turns out that  $Z_i = 2.32$  for all sequences  $x_i$ .

Figure 2.5 shows the relative-frequency and kernel-density estimates for our toy problem. Figure 2.5a shows the 18 forms in the English lexicon, and a relative-frequency estimate of the distribution over possible forms if each such entry from the English lexicon is counted as a single observation. Figure 2.5b shows the kernel density estimate. The unattested lexical forms now have non-zero probability, and furthermore the probability of both attested and unattested forms depends on how densely their neighborhoods in phonological space are occupied.

#### 2.11.4 Kernel density estimation and exemplar models

As is made explicit in Equation (2.27), computing the probability of an outcome using kernel density estimation (KDE) involves iterating explicitly over the entire set of observations  $\mathbf{y}$ . From a computational point of view, a distinctive property of KDE is that it requires the storage and access of complete datasets. In theoretical linguistics, psycholinguistics, and computational linguistics, models with this requirement are often called EXEMPLAR MODELS. Exemplar models have received considerable attention in these fields as candidate models of language acquisition and use. For example, Bailey and Hahn (2001) adapted the exemplar

model of Nosofsky (1986) to the problem of inducing probability distributions over possible lexical forms (Section 2.11.3). In syntax, the Data-Oriented Parsing model (Scha, 1990; Bod, 1992, 1998, 2006) is perhaps the best known formalized exemplar model. A point that cannot be over-emphasized is that the core substance of an exemplar model consists of (a) the representation of the space of possible exemplars, and (b) the metric of similarity between points in the exemplar representation space. In kernel density estimation, the choice of kernel constitutes the metric of similarity.

One of the common criticisms of exemplar-based models is on the grounds that it is psychologically implausible to imagine that all the exemplars in a language user's experience are retained, and are exhaustively accessed whenever the probability of a given point in the exemplar representation space is needed. We won't take sides on this issue here. However, in future chapters we do consider methods that do not have the requirement of exhaustive storage and recall of the dataset.

### 2.11.5 Evaluating estimated probability densities

You now have basic tools for estimating probability densities from data. Even after this brief introduction to density estimation it should be painfully obvious that there are many ways to estimate a density from a given dataset. From the standpoint of modeling linguistic cognition, this is in fact advantageous because different approaches to estimation encode different *learning biases*. This means that density estimation procedures as models of human language learning can be evaluated in terms of how closely they reflect the inferences made by language learners from exposure to finite data.

There are many times, however, when you will also be interested in evaluating how well a particular density estimate intrinsically encodes the data from which it is derived. Here we cover two approaches to this type of evaluation that enjoy wide currency: CLASSIFICATION ACCURACY and LIKELIHOOD.

#### Classification accuracy

For discrete random variables, the simplest method of evaluating the performance of a density estimate is CLASSIFICATION ACCURACY. Prerequisite to this notion is the notion of PREDICTION: for a given context (i.e. conditioning on some known information), what value of a discrete random variable  $X$  do I expect? Suppose that I have a probability density estimate based on some observations  $\mathbf{y}$  and I expect to see  $n$  new observations. My predictions for the observed outcomes in these new observations (before I actually see them) can be labeled  $\hat{y}_1, \dots, \hat{y}_n$ . My classification accuracy is simply the proportion of these predictions that turn out to be correct. For example, suppose that for the German grammatical function ordering data of Section 2.11.1 I expect to see three new observations. Without any further information, the most sensible thing for me to do is simply to predict  $\langle \mathbf{SB}, \mathbf{DO} \rangle$  all three times, since I estimated it to be the most likely outcome. If the actual outcomes are  $\langle \mathbf{SB}, \mathbf{DO} \rangle, \langle \mathbf{SB}, \mathbf{IO} \rangle, \langle \mathbf{SB}, \mathbf{DO} \rangle$ , then my classification accuracy is  $\frac{2}{3}$ .

When we obtain a dataset all at once, then a common practice for evaluating the classification accuracy of a density estimation technique for making predictions from that data is to remove a small part of the dataset and use the rest to construct the density estimate. We then evaluate the classification accuracy of the density estimate on the small part of the dataset that was removed. For example, in the German word order data we might draw 10% of the data (57 observations)—say 45 of  $\langle \mathbf{SB}, \mathbf{DO} \rangle$ , 7 of  $\langle \mathbf{SB}, \mathbf{IO} \rangle$ , 3 of  $\langle \mathbf{IO}, \mathbf{SB} \rangle$ , and two of  $\langle \mathbf{IO}, \mathbf{DO} \rangle$ . In the remaining 90% of the data,  $\langle \mathbf{SB}, \mathbf{DO} \rangle$  remains the most likely outcome, so we predict that for all 57 of the removed observations. Our classification accuracy is thus  $\frac{45}{57} \approx 0.79$ . This approach is called HELD-OUT EVALUATION.

Of course, held-out evaluation has some disadvantages as well, notably that the evaluation is based on only a small fraction of available data and hence will be noisy. Another widely used technique is CROSS-VALIDATION, in which we split our dataset into  $k$  equally sized portions. We then produce  $k$  different held-out evaluations of classification accuracy, each portion in turn serving as the held-out portion of the dataset, and average the classification accuracy across the folds. As a simple example, suppose we collected a corpus of  $\{\textit{night}, \textit{day}\}$  binomials, with 160 examples of *day and night* (**d**) and 140 examples of *night and day* (**n**). On 4-fold cross-validation, we might obtain the following outcomes:

| Fold  | # d held out | # n held out | Prediction | Accuracy |
|-------|--------------|--------------|------------|----------|
| 1     | 34           | 41           | <b>n</b>   | 0.45     |
| 2     | 40           | 35           | <b>d</b>   | 0.53     |
| 3     | 48           | 27           | <b>d</b>   | 0.36     |
| 4     | 38           | 37           | <b>d</b>   | 0.51     |
| Total |              |              |            | 0.46     |

If computing the predictions from observed data is fast, then the best kind of cross-validation is generally LEAVE-ONE-OUT cross-validation, where there are as many folds as there are observations.

### Likelihood-based evaluation

Classification accuracy is “brittle” in the sense that it discards a great deal of information from a density estimate. As a simple example, for the binomial  $\{\textit{night}, \textit{day}\}$  the best classification decision is **d** if  $\hat{P}(\mathbf{d}) > 0.5$ ; but  $\hat{P}(\mathbf{d}) = 0.51$  and  $\hat{P}(\mathbf{d}) = 0.99$  are very different distributions, and we’d like to distinguish the types of predictions they make. Furthermore, classification accuracy doesn’t even make sense in the continuous random variable setting. We can address both these problems, however, via the concept of LIKELIHOOD, which we briefly encountered in Section 2.4.1. The likelihood under a density estimate  $\hat{P}$  of a set of data  $\mathbf{y}$  is simply

$$P(\mathbf{y}|\hat{P}) \tag{2.28}$$

The likelihood is sometimes viewed as a function of a set of observations  $\mathbf{y}$ , and sometimes (see Chapter 4) viewed as a property of the estimate itself  $\hat{P}$ . If the observations  $y_i$  are assumed to be independent of one another, then we can rewrite the likelihood as

$$P(\mathbf{y}|\hat{P}) = \prod_{i=1}^n P(y_i|\hat{P}) \quad (2.29)$$

Since likelihoods can be very small, and datasets can be very large, explicitly computing the product in Equation (2.29) can lead to problems with computational underflow. This can be avoided by computing the LOG-LIKELIHOOD; in log-space you are extremely unlikely to have computational underflow or overflow problems. Since the log of a product is the sum of a log, you'll usually see log-likelihood computed as in Equation (2.30) below:

$$\log P(\mathbf{y}|\hat{P}) = \sum_{i=1}^n \log P(y_i|\hat{P}) \quad (2.30)$$

Likelihood can be evaluated with respect to the data used to estimate the density, with respect to held-out data, or using cross-validation. Evaluating log-likelihood from the data used to estimate the model is, however, a dangerous enterprise. This is illustrated in Figures 2.6 through 2.8, for the example from Section 2.11.2 of estimating F0 formant frequency through normal-kernel density estimation. Figure 2.6 illustrates the change in estimated probability density as a result of choosing different bandwidths. The narrower the bandwidth, the more of the probability mass is focused around the observations themselves. As a result, the log-likelihood of the data used to estimate the density increases monotonically as the bandwidth decreases (Figure 2.7). The likelihood as evaluated with six-fold cross-validation, on the other hand reaches a maximum at bandwidth  $b \approx 10\text{Hz}$  (Figure 2.8). The discrepancy between the shapes of the curves in Figures 2.7 and 2.8 for  $b < 10$ —and also of the generally much higher log-likelihoods in the former figure—reveals that the narrow-bandwidth density estimates are OVERFITTING—intuitively, they mimic the observed data too closely and generalize too little. The cross-validated likelihood reveals that the assessment of The ability of the narrow-bandwidth density estimate to closely mimic observed data is a kind of *complexity* of the estimation process. Finding a balance between complexity and generalization is a hallmark issue of statistical inference, and we will see this issue arise again in numerous contexts throughout the book.

## 2.12 Exercises

### Exercise 2.1: Conditional independence and set intersection<sup>†</sup>

Show that if  $A$  and  $B$  are conditionally independent given  $C$  and  $P(A|C) > 0$ , then  $P(B|A \cap C) = P(B|C)$ . **Hint:** one natural solution involves making use of the definition of conditional independence in two different ways.

### Exercise 2.2: Loss of conditional independence<sup>\*\*</sup>



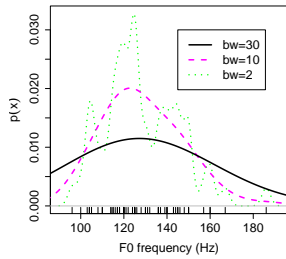


Figure 2.6: F0 formant kernel density estimates for normal kernels of different bandwidths.

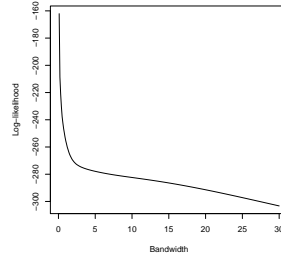


Figure 2.7: Overfitting when evaluating on training set: F0 formant log-likelihood increases unboundedly.

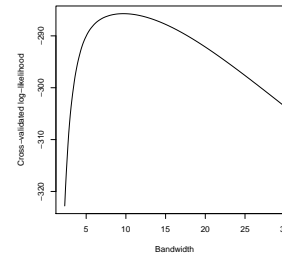


Figure 2.8: Cross-validated log-likelihood reveals the optimal bandwidth.

Give an example in words where two events  $A$  and  $B$  are conditionally independent given some state of knowledge  $C$ , but when another piece of knowledge  $D$  is learned,  $A$  and  $B$  lose conditional independence.

### Exercise 2.3: tea in *Wonderland*♣

1. You obtain infinitely many copies of the text *Alice in Wonderland* and decide to play a word game with it. You cut apart each page of each copy into individual letters, throw all the letters in a bag, shake the bag, and draw three letters at random from the bag. What is the probability that you will be able to spell **tea**?
2. Why did the problem specify that you obtained infinitely many copies of the text? Suppose that you obtained only one copy of the text? Would you have enough information to compute the probability of being able to spell **tea**? Why?

### Exercise 2.4: Bounds on probability density functions\*

- Discrete random variables are governed by probability mass functions, which are bounded below by zero and above by 1 (that is, every value a probability mass function must be at least zero and no more than 1). Why must a probability mass function be bounded above by 1?
- Continuous random variables are governed by probability density functions, which are bounded below by zero. What are probability density functions bounded above by? Why?

### Exercise 2.5: Probabilities in the constituent-order model♣

For the constituent-order example given in 2.8, let  $\gamma_1 = 0.6$ ,  $\gamma_2 = 0.4$ , and  $\gamma_3 = 0.3$ . Compute the probabilities of all six possible word orders.

**Exercise 2.6: Parameters of the constituent-order model\*\***

In the constituent-order example given in 2.8, I mentioned that we would like to interpret the probability of each parameter  $\gamma_i$  analogously to the success parameter of a single Bernoulli trial, so that the probability that  $S \prec O$  is  $\gamma_1$ , the probability that  $S \prec V$  is  $\gamma_2$ , and the probability that  $O \prec V$  is  $\gamma_3$ . Given the mass function  $F$  actually used in the example, is the probability that  $S \prec O$  actually  $\gamma_1$ ? Show your work.

**Exercise 2.7: Expectation of a uniform random variable\***

Prove mathematically that the expectation of a uniform random variable  $X$  on  $[a, b]$  is  $E(X) = \frac{b+a}{2}$ . (This involves a simple integration; consult A.5 if you need a refresher.)

**Exercise 2.8: Variance of a uniform random variable\*\***

Prove that the variance of a continuous, uniformly distributed random variable  $X$  on  $[a, b]$  is  $\frac{(b-a)^2}{12}$ .

**Exercise 2.9: Normal distributions\***

For adult female native speakers of American English, the distribution of first-formant frequencies for the vowel  $[\varepsilon]$  is reasonably well modeled as a normal distribution with mean 608Hz and standard deviation 77.5Hz. What is the probability that the first-formant frequency of an utterance of  $[\varepsilon]$  for a randomly selected adult female native speaker of American English will be between 555Hz and 697Hz?

**Exercise 2.10: Assessing optimal kernel bandwidth through cross-validation<sup>†</sup>**

Use leave-one-out cross-validation to calculate the cross-validated likelihood of kernel density estimates (using a normal kernel) of adult male speaker  $[a]$  and  $[i]$  F2 formants from the Peterson and Barney dataset. Plot the cross-validated likelihood as a function of kernel bandwidth. Are the bandwidths that work best for  $[a]$  and  $[i]$  similar to each other? Explain your results.

**Exercise 2.11: Kernels in discrete linguistic spaces<sup>‡</sup>**

Construct your own kernel over the space of 27 CVC lexical forms used in Section 2.11.3. With the 18 attested forms listed in that section, use leave-one-out cross-validation to compute the cross-validated likelihood of your kernel. Does it do better or worse than the original kernel?

**Exercise 2.12: Exemplar-based model of phonotactic knowledge<sup>‡</sup>**

The file `syllCounts` contains phonetic representations of attested word-initial syllables in disyllabic words of English, based on the CELEX lexical database, together with their frequencies of occurrence. Find a partner in your class. Each of you should independently construct a kernel over the phonetic representations of these word-initial syllables. Using ten-fold cross-validation, compute the cross-validated likelihood of kernel density estimates using each of your kernels. Compare the results, noting which syllables each kernel did better on.

Now join forces and try to construct a third kernel which combines the best qualities of your two kernels, and (hopefully) has a higher cross-validated likelihood than either one.

# Chapter 3

## Multivariate Probability

### 3.1 Joint probability mass and density functions

Recall that a basic probability distribution is defined over a random variable, and a random variable maps from the sample space to the real numbers. What about when you are interested in the outcome of an event that is not naturally characterizable as a single real-valued number, such as the two formants of a vowel?

The answer is simple: probability mass and density functions can be generalized over multiple random variables at once. If all the random variables are discrete, then they are governed by a JOINT PROBABILITY MASS FUNCTION; if all the random variables are continuous, then they are governed by a JOINT PROBABILITY DENSITY FUNCTION. There are many things we'll have to say about the joint distribution of collections of random variables which hold equally whether the random variables are discrete, continuous, or a mix of both.<sup>1</sup> In these cases we will simply use the term “joint density” with the implicit understanding that in some cases it is a probability mass function.

Notationally, for random variables  $X_1, X_2, \dots, X_N$ , the joint density is written as

$$p(X_1 = x_1, X_2 = x_2, \dots, X_N = x_n) \tag{3.1}$$

or simply

$$p(x_1, x_2, \dots, x_n) \tag{3.2}$$

for short.

---

<sup>1</sup>If some of the random variables are discrete and others are continuous, then technically it is a probability density function rather than a probability mass function that they follow; but whenever one is required to compute the total probability contained in some part of the range of the joint density, one must sum on the discrete dimensions and integrate on the continuous dimensions.

### 3.1.1 Joint cumulative distribution functions

For a single random variable, the cumulative distribution function is used to indicate the probability of the outcome falling on a segment of the real number line. For a collection of  $N$  random variables  $X_1, \dots, X_N$  (or density), the analogous notion is the JOINT CUMULATIVE DISTRIBUTION FUNCTION, which is defined with respect to regions of  $N$ -dimensional space. The joint cumulative distribution function, which is sometimes notated as  $F(x_1, \dots, x_n)$ , is defined as the probability of the set of random variables all falling at or below the specified values of  $X_i$ :<sup>2</sup>

$$F(x_1, \dots, x_n) \stackrel{\text{def}}{=} P(X_1 \leq x_1, \dots, X_N \leq x_n)$$

The natural thing to do is to use the joint cpd to describe the probabilities of rectangular volumes. For example, suppose  $X$  is the  $f_1$  formant and  $Y$  is the  $f_2$  formant of a given utterance of a vowel. The probability that the vowel will lie in the region  $480\text{Hz} \leq f_1 \leq 530\text{Hz}$ ,  $940\text{Hz} \leq f_2 \leq 1020\text{Hz}$  is given below:

$$P(480\text{Hz} \leq f_1 \leq 530\text{Hz}, 940\text{Hz} \leq f_2 \leq 1020\text{Hz}) = \\ F(530\text{Hz}, 1020\text{Hz}) - F(530\text{Hz}, 940\text{Hz}) - F(480\text{Hz}, 1020\text{Hz}) + F(480\text{Hz}, 940\text{Hz})$$

and visualized in Figure 3.1 using the code below.

## 3.2 Marginalization

Often we have direct access to a joint density function but we are more interested in the probability of an outcome of a subset of the random variables in the joint density. Obtaining this probability is called MARGINALIZATION, and it involves taking a weighted sum<sup>3</sup> over the possible outcomes of the random variables that are not of interest. For two variables  $X, Y$ :

---

<sup>2</sup>Technically, the definition of the multivariate cumulative distribution function is

$$F(x_1, \dots, x_n) \stackrel{\text{def}}{=} P(X_1 \leq x_1, \dots, X_N \leq x_n) = \sum_{\vec{x} \leq \langle x_1, \dots, x_n \rangle} p(\vec{x}) \quad [\text{Discrete}] \quad (3.3)$$

$$F(x_1, \dots, x_n) \stackrel{\text{def}}{=} P(X_1 \leq x_1, \dots, X_N \leq x_n) = \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_n} p(\vec{x}) dx_n \dots dx_1 \quad [\text{Continuous}] \quad (3.4)$$

<sup>3</sup>or integral in the continuous case

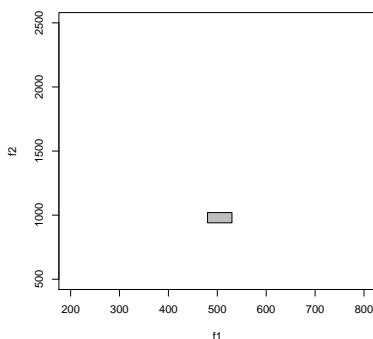


Figure 3.1: The probability of the formants of a vowel landing in the grey rectangle can be calculated using the joint cumulative distribution function.

$$\begin{aligned}
 P(X = x) &= \sum_y P(x, y) \\
 &= \sum_y P(X = x|Y = y)P(y)
 \end{aligned}$$

In this case  $P(X)$  is often called a *marginal density* and the process of calculating it from the joint density  $P(X, Y)$  is known as *marginalization*.

As an example, consider once again the historical English example of Section 2.4. We can now recognize the table in I as giving the joint density over two binary-valued random variables: the position of the object with respect to the verb, which we can denote as  $X$ , and the pronominality of the object NP, which we can denote as  $Y$ . From the joint density given in that section we can calculate the marginal density of  $X$ :

$$P(X = x) = \begin{cases} 0.224 + 0.655 = 0.879 & x = \mathbf{Preverbal} \\ 0.014 + 0.107 = 0.121 & x = \mathbf{Postverbal} \end{cases} \quad (3.5)$$

Additionally, if you now look at the old English example of Section 2.4.1 and how we calculated the denominator of Equation 2.7, you will see that it involved marginalization over the animacy of the object NP. Repeating Bayes' rule for reference:

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

It is very common to need to explicitly marginalize over  $A$  to obtain the marginal probability for  $B$  in the computation of the denominator of the right-hand side.

### 3.3 Linearity of expectation, covariance, correlation, and variance sums of random variables

#### 3.3.1 Linearity of the expectation

Linearity of the expectation is an extremely important property and can be expressed in two parts. First, if you *rescale* a random variable, its expectation rescales in the exact same way. Mathematically, if  $Y = a + bX$ , then  $E(Y) = a + bE(X)$ .

Second, the expectation of the sum of random variables is the sum of the expectations. That is, if  $Y = \sum_i X_i$ , then  $E(Y) = \sum_i E(X_i)$ . This holds regardless of any conditional dependencies that hold among the  $X_i$ .

We can put together these two pieces to express the expectation of a linear combination of random variables. If  $Y = a + \sum_i b_i X_i$ , then

$$E(Y) = a + \sum_i b_i E(X_i) \tag{3.6}$$

This is incredibly convenient. We'll demonstrate this convenience when we introduce the binomial distribution in Section 3.4.

#### 3.3.2 Covariance

The COVARIANCE between two random variables  $X$  and  $Y$  is a measure of how tightly the outcomes of  $X$  and  $Y$  tend to pattern together. It is defined as follows:

$$\text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))]$$

When the covariance is positive,  $X$  tends to be high when  $Y$  is high, and vice versa; when the covariance is negative,  $X$  tends to be high when  $Y$  is low, and vice versa.

As a simple example of covariance we'll return once again to the Old English example of Section 2.4; we repeat the joint density for this example below, with the marginal densities in the row and column margins:

|     |                          | Coding for $Y$          |                    |       |
|-----|--------------------------|-------------------------|--------------------|-------|
|     |                          | 0                       | 1                  |       |
| (1) | Coding for $X$           | <b>Pronoun</b>          | <b>Not Pronoun</b> |       |
|     | 0                        | Object <b>Preverbal</b> | 0.224              | 0.655 |
| 1   | Object <b>Postverbal</b> | 0.014                   | 0.107              | .121  |
|     |                          | .238                    | .762               |       |

We can compute the covariance by treating each of  $X$  and  $Y$  as a Bernoulli random variable, using arbitrary codings of 1 for **Postverbal** and **Not Pronoun**, and 0 for **Preverbal** and

**Pronoun.** As a result, we have  $E(X) = 0.121$ ,  $E(Y) = 0.762$ . The covariance between the two can then be computed as follows:

$$\begin{aligned}
 & (0 - 0.121) \times (0 - .762) \times .224 && \text{(for X=0,Y=0)} \\
 & +(1 - 0.121) \times (0 - .762) \times 0.014 && \text{(for X=1,Y=0)} \\
 & +(0 - 0.121) \times (1 - .762) \times 0.655 && \text{(for X=0,Y=1)} \\
 & +(1 - 0.121) \times (1 - .762) \times 0.107 && \text{(for X=1,Y=1)} \\
 & =0.014798
 \end{aligned}$$

If  $X$  and  $Y$  are conditionally independent given our state of knowledge, then  $\text{Cov}(X, Y)$  is zero (Exercise 3.1 asks you to prove this).

### 3.3.3 Covariance and scaling random variables

What happens to  $\text{Cov}(X, Y)$  when you scale  $X$ ? Let  $Z = a + bX$ . It turns out that the covariance with  $Y$  increases by  $b$  (Exercise 3.3 asks you to prove this):

$$\text{Cov}(Z, Y) = b\text{Cov}(X, Y)$$

As an important consequence of this, rescaling a random variable by  $Z = a + bX$  rescales its variance by  $b^2$ :  $\text{Var}(Z) = b^2\text{Var}(X)$  (see Exercise 3.2).

### 3.3.4 Correlation

We just saw that the covariance of word length with frequency was much higher than with log frequency. However, the covariance cannot be compared directly across different pairs of random variables, because we also saw that random variables on different scales (e.g., those with larger versus smaller ranges) have different covariances due to the scale. For this reason, it is common to use the CORRELATION  $\rho$  as a standardized form of covariance:

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

If  $X$  and  $Y$  are independent, then their covariance (and hence correlation) is zero.

### 3.3.5 Variance of the sum of random variables

It is quite often useful to understand how the variance of a sum of random variables is dependent on their joint distribution. Let  $Z = X_1 + \dots + X_n$ . Then

$$\text{Var}(Z) = \sum_{i=1}^n \text{Var}(X_i) + \sum_{i \neq j} \text{Cov}(X_i, X_j) \tag{3.7}$$



Since the covariance between conditionally independent random variables is zero, it follows that the variance of the sum of pairwise independent random variables is the sum of their variances.

## 3.4 The binomial distribution

We're now in a position to introduce one of the most important probability distributions for linguistics, the BINOMIAL DISTRIBUTION. The binomial distribution family is characterized by two parameters,  $n$  and  $\pi$ , and a binomially distributed random variable  $Y$  is defined as the sum of  $n$  identical, independently distributed (i.i.d.) Bernoulli random variables, each with parameter  $\pi$ .

For example, it is intuitively obvious that the mean of a binomially distributed r.v.  $Y$  with parameters  $n$  and  $\pi$  is  $\pi n$ . However, it takes some work to show this explicitly by summing over the possible outcomes of  $Y$  and their probabilities. On the other hand,  $Y$  can be re-expressed as the sum of  $n$  BERNOULLI RANDOM VARIABLES  $X_i$ . The resulting probability density function is, for  $k = 0, 1, \dots, n$ :<sup>4</sup>

$$P(Y = k) = \binom{n}{k} \pi^k (1 - \pi)^{n-k} \quad (3.8)$$

We'll also illustrate the utility of the linearity of expectation by deriving the expectation of  $Y$ . The mean of each  $X_i$  is trivially  $\pi$ , so we have:

$$E(Y) = \sum_i^n E(X_i) \quad (3.9)$$

$$= \sum_i^n \pi = \pi n \quad (3.10)$$

which makes intuitive sense.

Finally, since a binomial random variable is the sum of  $n$  mutually independent Bernoulli random variables and the variance of a Bernoulli random variable is  $\pi(1 - \pi)$ , the variance of a binomial random variable is  $n\pi(1 - \pi)$ .

### 3.4.1 The multinomial distribution

The MULTINOMIAL DISTRIBUTION is the generalization of the binomial distribution to  $r \geq 2$  possible outcomes. (It can also be seen as the generalization of the distribution over multinomial trials introduced in Section 2.5.2 to the case of  $n \geq 1$  trials.) The  $r$ -class multinomial is

---

<sup>4</sup>Note that  $\binom{n}{k}$  is pronounced “ $n$  choose  $k$ ”, and is defined as  $\frac{n!}{k!(n-k)!}$ . In turn,  $n!$  is pronounced “ $n$  factorial”, and is defined as  $n \times (n - 1) \times \dots \times 1$  for  $n = 1, 2, \dots$ , and as 1 for  $n = 0$ .

a sequence of  $r$  random variables  $X_1, \dots, X_r$  whose joint distribution is characterized by  $r$  parameters: a size parameter  $n$  denoting the number of trials, and  $r-1$  parameters  $\pi_1, \dots, \pi_{r-1}$ , where  $\pi_i$  denotes the probability that the outcome of a single trial will fall into the  $i$ -th class. (The probability that a single trial will fall into the  $r$ -th class is  $\pi_r \stackrel{\text{def}}{=} 1 - \sum_{i=1}^{r-1} \pi_i$ , but this is not a real parameter of the family because it's completely determined by the other parameters.) The (joint) probability mass function of the multinomial looks like this:

$$P(X_1 = n_1, \dots, X_r = n_r) = \binom{n}{n_1 \dots n_r} \prod_{i=1}^r \pi_i \quad (3.11)$$

where  $n_i$  is the number of trials that fell into the  $r$ -th class, and  $\binom{n}{n_1 \dots n_r} = \frac{n!}{n_1! \dots n_r!}$ .

### 3.5 Multivariate normal distributions

Finally, we turn to the MULTIVARIATE NORMAL DISTRIBUTION. Recall that the univariate normal distribution placed a probability density over outcomes of a single continuous random variable  $X$  that was characterized by two parameters—mean  $\mu$  and variance  $\sigma^2$ . The multivariate normal distribution in  $N$  dimensions, in contrast, places a joint probability density on  $N$  real-valued random variables  $X_1, \dots, X_N$ , and is characterized by two sets of parameters: (1) a mean vector  $\mu$  of length  $N$ , and (2) a symmetric COVARIANCE MATRIX (or variance-covariance matrix)  $\Sigma$  in which the entry in the  $i$ -th row and  $j$ -th column expresses the covariance between  $X_i$  and  $X_j$ . Since the covariance of a random variable with itself is its variance, the diagonal entries of  $\Sigma$  are the variances of the individual  $X_i$  and must be non-negative.

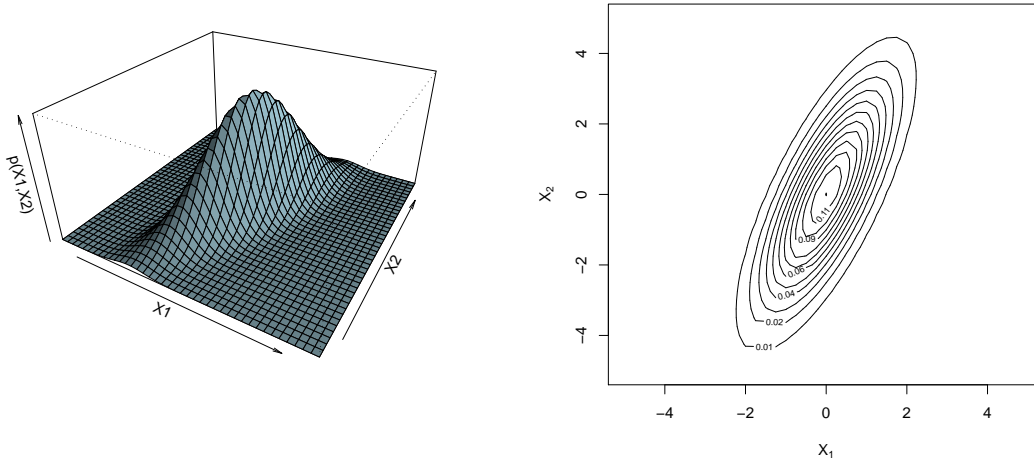
The probability density function for the multivariate normal distribution is most easily expressed using matrix notation (Section A.8); the symbol  $\mathbf{x}$  stands for the vector  $\langle x_1, \dots, x_n \rangle$ :

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi|\Sigma|)^N}} \exp \left[ -\frac{(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)}{2} \right] \quad (3.12)$$

For example, a bivariate normal distribution ( $N = 2$ ) over random variables  $X_1$  and  $X_2$  has two means  $\mu_1, \mu_2$ , and the covariance matrix contains two variance terms (one for  $X_1$  and one for  $X_2$ ), and one *covariance term* showing the correlation between  $X_1$  and  $X_2$ . The covariance matrix would look like  $\begin{pmatrix} \sigma_{11}^2 & \sigma_{12}^2 \\ \sigma_{12}^2 & \sigma_{22}^2 \end{pmatrix}$ . Once again, the terms  $\sigma_{11}^2$  and  $\sigma_{22}^2$  are simply the variances of  $X_1$  and  $X_2$  respectively (the subscripts appear doubled for notational consistency). The term  $\sigma_{12}^2$  is the covariance between the two axes.<sup>5</sup> Figure 3.2 visualizes a

<sup>5</sup>The probability density function works out to be

$$p(x_1, x_2) = \frac{1}{2\pi\sqrt{\sigma_{11}^2\sigma_{22}^2 - \sigma_{12}^4}} \exp \left[ \frac{(x_1 - \mu_1)^2\sigma_{22}^2 - 2(x_1 - \mu_1)(x_2 - \mu_2)\sigma_{12}^2 + (x_2 - \mu_2)^2\sigma_{11}^2}{\sigma_{11}^2\sigma_{22}^2 - \sigma_{12}^4} \right]$$



(a) Perspective plot

(b) Contour plot

Figure 3.2: Visualizing the multivariate normal distribution

bivariate normal distribution with  $\mu = (0, 0)$  and  $\Sigma = \begin{pmatrix} 1 & 1.5 \\ 1.5 & 4 \end{pmatrix}$ . Because the variance is larger in the  $X_2$  axis, probability density falls off more rapidly along the  $X_1$  axis. Also note that the major axis of the ellipses of constant probability in Figure 3.2b does not lie right on the  $X_2$  axis, but rather is at an angle reflecting the positive covariance.

The multivariate normal distribution is very useful in modeling multivariate data such as the distribution of multiple formant frequencies in vowel production. As an example, Figure 3.3 shows how a large number of raw recordings of five vowels in American English can be summarized by five “characteristic ellipses”, one for each vowel. The center of each ellipse is placed at the empirical mean for the vowel, and the shape of the ellipse reflects the empirical covariance matrix for that vowel.

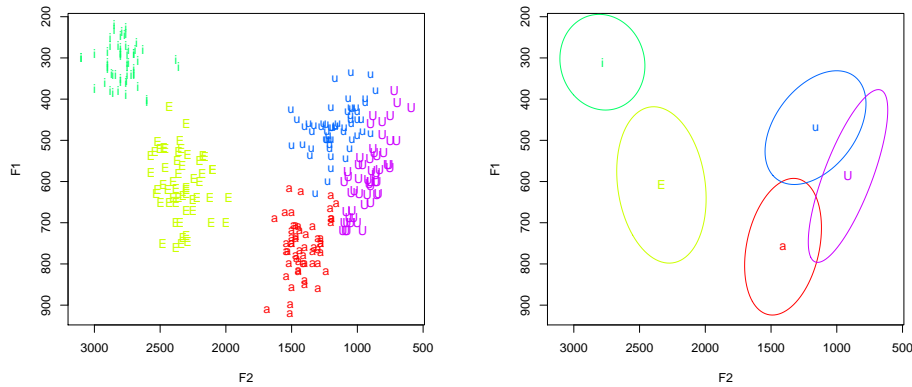
In addition, multivariate normal distributions plays an important role in almost all hierarchical models, covered starting in Chapter 8.

### 3.6 The central limit theorem

The CENTRAL LIMIT THEOREM is a powerful result from probability theory that states that the sum of a large quantity of i.i.d. random variables will have an approximately normal distribution, *regardless of the distribution of the individual random variables*. More formally, suppose that we have  $n$  i.i.d. random variables  $X_i$ , with  $Y = X_1 + \dots + X_n$ . From linearity of the variance, we know that  $Y$ 's mean is  $\mu_Y = nE[X_i]$ , and its variance is  $\sigma_Y^2 = n\sigma_{X_i}^2$ . The

---

Note that if  $\sigma_{11}$  is much larger than  $\sigma_{22}$ , then  $x_2 - \mu_2$  will be more important than  $x_1 - \mu_1$  in the exponential. This reflects the fact that if the variance is much larger on the  $X_1$  axis than on the  $X_2$  axis, a fixed amount of deviation from the mean is much less probable along the  $x_2$  axis.



(a) Raw recordings of five vowels by adult female speakers (b) Representation as multivariate normal distributions. The character is placed at the empirical mean for each vowel, and the covariance structure of each vowel is represented by an equiprobability ellipse

Figure 3.3: F1 and F2 formant frequency representations using multivariate normal distributions, based on the data of Peterson and Barney (1952)

central limit theorem states that as the number  $n$  of random variables grows, the distribution of the random variable  $Z = \frac{Y - \mu_Y}{\sigma_Y}$  approaches that of a standard normal random variable.

The central limit theorem traditionally serves as the basis for using the normal distribution to model the outcome of a complex process with many underlying contributing factors. Exercise 3.11 explores a simple example illustrating the truth of the theorem, showing how a binomial distribution with large  $n$  can be approximated by the normal distribution.

### 3.7 Exercises

#### Exercise 3.1: Covariance of conditionally independent random variables.

Use linearity of the expectation to prove that if two random variables  $X$  and  $Y$  are conditionally independent given your state of knowledge, then  $\text{Cov}(X, Y) = 0$  under this state of knowledge. (**Hint:** you can rewrite  $\sum_{x,y} Xp(X = x)Yp(Y = y)$  as  $\sum_x Xp(X = x) \sum_y Yp(Y = y)$ , since  $X$  and  $p(X = x)$  are constant with respect to  $y$ .)

#### Exercise 3.2: ♣

- What is the covariance of a random variable  $X$  with itself?
- Now show that if you rescale a random variable  $X$  by defining  $Z = a + bX$ , then  $\text{Var}(Z) = b^2\text{Var}(X)$ .

### Exercise 3.3

Show that if you rescale  $X$  as  $Z = a + bX$ , then  $\text{Cov}(Z, Y) = b\text{Cov}(X, Y)$ .

### Exercise 3.4

Prove Equation 3.7—that is, that  $\text{Var}(X_1 + \dots + X_n) = \sum_{i=1}^n \text{Var}(X_i) + \sum_{i \neq j} \text{Cov}(X_i, X_j)$ .

### Exercise 3.5

Let's return to coin flipping, but use a different process to generate a sequence of coin flips. Suppose I start flipping a coin with success parameter  $\pi$ , and every time it comes up tails I keep on flipping, but the first time it comes up heads I stop. The random variable of interest is the length of the sequence of coin flips. The GEOMETRIC DISTRIBUTION characterizes the probability density on this random variable. The probability mass function of the geometric distribution has the form

$$P(X = k) = (1 - \pi)^a \pi^b, k \in \{1, 2, \dots\}$$

for some choice of  $a$  and  $b$ . Complete the specification of the distribution (i.e., say what  $a$  and  $b$  are) and justify it.

### Exercise 3.6

The file `brown-counts-lengths-nsyll` contains the following properties for each word type found in the parsed Brown corpus:

- The token frequency of the word;
- The length of the word in letters;
- The number of syllables in the word, as determined by the CMU Pronouncing dictionary (<http://www.speech.cs.cmu.edu/cgi-bin/cmudict>).

Plot histograms of the number of syllables for word, over (a) **word types** and (b) **word tokens**. Which of the histograms looks more binomially-distributed? Which looks more geometrically-distributed? Try to find a good fit (by eyeball assessment) to each of the histograms by choosing binomial or geometric parameters that match the data as well as you can.

### Exercise 3.7

The NEGATIVE BINOMIAL distribution is a generalization of the geometric distribution in which you are interested in how many coin flips you can make before you achieve a total of  $r$  successes (where the successes are included in the total number of flips). The distribution is characterized by two parameters: the required number of successes  $r$ , and the probability  $p$  of success on any given coin flip. (The geometric distribution is a negative binomial distribution for which  $r = 1$ .) If the total number of coin flips obtained in a given trial is  $k$ , then the probability mass function for a negative binomial distribution with parameters  $p, r$  has the form

$$P(X = k; r, p) = \binom{a}{b} (1-p)^c p^d, k \in \{r, r+1, \dots\}$$

for some choice of  $a, b, c, d$ . Complete the specification of the distribution (i.e., say what  $a, b, c, d$  are) and justify it.

### Exercise 3.8: Linearity of expectation

You put two coins in a pouch; one coin is weighted such that it lands heads  $\frac{5}{6}$  of the time when it's flipped, and the other coin is weighted such that it lands heads  $\frac{1}{3}$  of the time when it's flipped. You shake the pouch, choose one of the coins from it at random and flip it twice. Write out both the marginal density for the outcome of the first flip and the joint density for the outcome of the two coin flips. Define the random variable  $X$  as the number of heads resulting from the two coin flips. Use linearity of the expectation to compute  $E(X)$ . Then compute  $E(X)$  directly from the joint density to confirm that linearity of the expectation holds.

### Exercise 3.9

Explain why rescaling a random variable by  $Z = a + bX$  changes the variance by a factor of  $b^2$ , so that  $\text{Var}(Z) = b^2 \text{Var}(X)$ . (See Section 3.3.3.)

### Exercise 3.10

You are planning on conducting a word recognition study using the lexical-decision paradigm, in which a participant is presented a letter sequence on a computer screen and then presses a key on the keyboard as soon as she recognizes it as either a word (the key F) or a non-word (the key J). The distribution of measured response times for non-words in this study is the sum of two independent random variables:  $X$ , the elapsed time from the appearance of the letter string on the screen to the participant's successful pressing of a key; and  $Y$ , the time elapsed between the pressing of the key and the successful recording of the key press by the computer (this distribution is governed by the polling rate and reliability of the keyboard). Suppose that  $X$  has mean 600 and standard deviation 80, and  $Y$  has mean 15 and standard deviation 9 (all measured in milliseconds). What are the mean and standard deviation of recorded reaction times ( $X + Y$ )?

### Exercise 3.11

Test the validity of the central limit theorem. Choose your own probability distribution, generate  $n$  i.i.d. random variables, add them together repeatedly, and standardize them (subtract out the mean and divide by the standard deviation). Use these multiple trials to generate estimated probability density and cumulative distribution functions. Compare these to the density and cumulative distribution function of the standard normal distribution. Do this for at least (a) the uniform and (b) the Bernoulli distribution. You're also welcome to use other distributions or invent your own.

# Chapter 4

## Parameter Estimation

Thus far we have concerned ourselves primarily with *probability theory*: what events may occur with what probabilities, given a model family and choices for the parameters. This is useful only in the case where we know the precise model family and parameter values for the situation of interest. But this is the exception, not the rule, for both scientific inquiry and human learning & inference. Most of the time, we are in the situation of processing data whose generative source we are uncertain about. In Chapter 2 we briefly covered elementary density estimation, using relative-frequency estimation, histograms and kernel density estimation. In this chapter we delve more deeply into the theory of probability density estimation, focusing on inference within parametric families of probability distributions (see discussion in Section 2.11.2). We start with some important properties of estimators, then turn to basic frequentist parameter estimation (maximum-likelihood estimation and corrections for bias), and finally basic Bayesian parameter estimation.

### 4.1 Introduction

Consider the situation of the first exposure of a native speaker of American English to an English variety with which she has no experience (e.g., Singaporean English), and the problem of inferring the probability of use of active versus passive voice in this variety with a simple transitive verb such as *hit*:

- (1) The ball hit the window. (Active)
- (2) The window was hit by the ball. (Passive)

There is ample evidence that this probability is contingent on a number of features of the utterance and discourse context (e.g., Weiner and Labov, 1983), and in Chapter 6 we cover how to construct such richer models, but for the moment we simplify the problem by assuming that active/passive variation can be modeled with a binomial distribution (Section 3.4) with parameter  $\pi$  characterizing the probability that a given potentially transitive clause eligible

for passivization will in fact be realized as a passive.<sup>1</sup> The question faced by the native American English speaker is thus, what inferences should we make about  $\pi$  on the basis of limited exposure to the new variety? This is the problem of PARAMETER ESTIMATION, and it is a central part of statistical inference. There are many different techniques for parameter estimation; any given technique is called an ESTIMATOR, which is applied to a set of data to construct an estimate. Let us briefly consider two simple estimators for our example.

**Estimator 1.** Suppose that our American English speaker has been exposed to  $n$  transitive sentences of the variety, and  $m$  of them have been realized in the passive voice in eligible clauses. A natural estimate of the binomial parameter  $\pi$  would be  $m/n$ . Because  $m/n$  is the relative frequency of the passive voice, this is known as the RELATIVE FREQUENCY ESTIMATE (RFE; see Section 2.11.1). In addition to being intuitive, we will see in Section 4.3.1 that the RFE can be derived from deep and general principles of optimality in estimation procedures. However, RFE also has weaknesses. For instance, it makes no use of the speaker's knowledge of her native English variety. In addition, when  $n$  is small, the RFE is unreliable: imagine, for example, trying to estimate  $\pi$  from only two or three sentences from the new variety.

**Estimator 2.** Our speaker presumably knows the probability of a passive in American English; call this probability  $q$ . An extremely simple estimate of  $\pi$  would be to ignore all new evidence and set  $\pi = q$ , regardless of how much data she has on the new variety. Although this option may not be as intuitive as Estimator 1, it has certain advantages: it is extremely reliable and, if the new variety is not too different from American English, reasonably accurate as well. On the other hand, once the speaker has had considerable exposure to the new variety, this approach will almost certainly be inferior to relative frequency estimation. (See Exercise to be included with this chapter.)

In light of this example, Section 4.2 describes how to assess the quality of an estimator in conceptually intuitive yet mathematically precise terms. In Section 4.3, we cover FREQUENTIST approaches to parameter estimation, which involve procedures for constructing point estimates of parameters. In particular we focus on maximum-likelihood estimation and close variants, which for multinomial data turns out to be equivalent to Estimator 1 above. In Section 4.4, we cover BAYESIAN approaches to parameter estimation, which involve placing probability distributions over the range of possible parameter values. The Bayesian estimation technique we will cover can be thought of as intermediate between Estimators 1 and 2.

## 4.2 Desirable properties for estimators

In this section we briefly cover three key properties of any estimator, and discuss the desirability of these properties.

---

<sup>1</sup>By this probability we implicitly conditionalize on the use of a transitive verb that is eligible for passivization, excluding intransitives and also unpassivizable verbs such as *weigh*.



### 4.2.1 Consistency

An estimator is CONSISTENT if the estimate  $\hat{\theta}$  it constructs is guaranteed to converge to the true parameter value  $\theta$  as the quantity of data to which it is applied increases. Figure 4.1 demonstrates that Estimator 1 in our example is consistent: as the sample size increases, the probability that the relative-frequency estimate  $\hat{\pi}$  falls into a narrow band around the true parameter  $\pi$  grows asymptotically toward 1 (this behavior can also be proved rigorously; see Section 4.3.1). Estimator 2, on the other hand, is not consistent (so long as the American English parameter  $q$  differs from  $\pi$ ), because it ignores the data completely. Consistency is nearly always a desirable property for a statistical estimator.

### 4.2.2 Bias

If we view the collection (or *sampling*) of data from which to estimate a population parameter as a stochastic process, then the parameter estimate  $\hat{\theta}_\eta$  resulting from applying a pre-determined estimator  $\eta$  to the resulting data can be viewed as a continuous random variable (Section 3.1). As with any random variable, we can take its expectation. In general, it is intuitively desirable that the expected value of the estimate be equal (or at least close) to the true parameter value  $\theta$ , but this will not always be the case. The BIAS of an estimator  $\eta$  is defined as the deviation of the expectation from the true value:  $E[\hat{\theta}_\eta] - \theta$ . All else being equal, the smaller the bias in an estimator the more preferable. An estimator for which the bias is zero—that is,  $E[\hat{\theta}_\eta] = \theta$ —is called UNBIASED.

Is Estimator 1 in our passive-voice example biased? The relative-frequency estimate  $\hat{\pi}$  is  $\frac{m}{n}$ , so  $E[\hat{\pi}] = E[\frac{m}{n}]$ . Since  $n$  is fixed, we can move it outside of the expectation (see linearity of the expectation in Section 3.3.1) to get

$$E[\hat{\pi}] = \frac{1}{n}E[m]$$

But  $m$  is just the number of passive-voice utterances heard, and since  $m$  is binomially distributed,  $E[m] = \pi n$ . This means that

$$\begin{aligned} E[\hat{\pi}] &= \frac{1}{n}\pi n \\ &= \pi \end{aligned}$$

So Estimator 1 is unbiased. Estimator 2, on the other hand, has bias  $q - \pi$ .

### 4.2.3 Variance (and efficiency)

Suppose that our speaker has decided to use Estimator 1 to estimate the probability  $\pi$  of a passive, and has been exposed to  $n$  utterances. The intuition is extremely strong that she should use *all*  $n$  utterances to form her relative-frequency estimate  $\hat{\pi}$ , rather than, say, using

only the first  $n/2$ . But why is this the case? Regardless of how many utterances she uses with Estimator 1, her estimate will be unbiased (think about this carefully if you are not immediately convinced). But our intuitions suggest that an estimate using less data is less reliable: it is likely to vary more dramatically due to pure freaks of chance.

It is useful to quantify this notion of reliability using a natural statistical metric: the VARIANCE of the estimator,  $\text{Var}(\hat{\theta})$  (Section 4.2.3). All else being equal, an estimator with smaller variance is preferable to one with greater variance. This idea, combined with a bit more simple algebra, quantitatively explains the intuition that more data is better for Estimator 1:

$$\begin{aligned}\text{Var}(\hat{\pi}) &= \text{Var}\left(\frac{m}{n}\right) \\ &= \frac{1}{n^2}\text{Var}(m) \quad (\text{From scaling a random variable, Section 3.3.3})\end{aligned}$$

Since  $m$  is binomially distributed, and the variance of the binomial distribution is  $n\pi(1 - \pi)$  (Section 3.4), so we have

$$\text{Var}(\hat{\pi}) = \frac{\pi(1 - \pi)}{n} \tag{4.1}$$

So variance is inversely proportional to the sample size  $n$ , which means that relative frequency estimation is more reliable when used with larger samples, consistent with intuition.

It is almost always the case that each of bias and variance comes at the cost of the other. This leads to what is sometimes called BIAS-VARIANCE TRADEOFF: one's choice of estimator may depend on the relative importance of expected accuracy versus reliability in the task at hand. The bias-variance tradeoff is very clear in our example. Estimator 1 is unbiased, but has variance that can be quite high when samples size  $n$  is small. Estimator 2 is biased, but it has zero variance. Which of the two estimators is preferable is likely to depend on the sample size. If our speaker anticipates that she will have very few examples of transitive sentences in the new English variety to go on, and also anticipates that the new variety will not be hugely different from American English, she may well prefer (and with good reason) the small bias of Estimator 2 to the large variance of Estimator 1. The lower-variance of two estimators is called the more EFFICIENT estimator, and the EFFICIENCY of one estimator  $\eta_1$  relative to another estimator  $\eta_2$  is the ratio of their variances,  $\text{Var}(\hat{\theta}_{\eta_1})/\text{Var}(\hat{\theta}_{\eta_2})$ .

### 4.3 Frequentist parameter estimation and prediction

We have just covered a simple example of parameter estimation and discussed key properties of estimators, but the estimators we covered were (while intuitive) given no theoretical underpinning. In the remainder of this chapter, we will cover a few major mathematically motivated estimation techniques of general utility. This section covers FREQUENTIST estimation techniques. In frequentist statistics, an estimator gives a point estimate for the parameter(s)

of interest, and estimators are preferred or dispreferred on the basis of their general behavior, notably with respect to the properties of consistency, bias, and variance discussed in Section 4.2. We start with the most widely-used estimation technique, MAXIMUM-LIKELIHOOD ESTIMATION.

### 4.3.1 Maximum Likelihood Estimation

We encountered the notion of the LIKELIHOOD in Chapter 2, a basic measure of the quality of a set of predictions with respect to observed data. In the context of parameter estimation, the likelihood is naturally viewed as a function of the parameters  $\theta$  to be estimated, and is defined as in Equation (2.28)—the joint probability of a set of observations, conditioned on a choice for  $\theta$ —repeated here:

$$\text{Lik}(\theta) \equiv P(y_1, \dots, y_n | \theta)$$

Since good predictions are better, a natural approach to parameter estimation is to choose the set of parameter values that yields the best predictions—that is, the parameter that *maximizes the likelihood* of the observed data. This value is called the MAXIMUM LIKELIHOOD ESTIMATE (MLE), defined formally as:<sup>2</sup>

$$\hat{\theta}_{MLE} \stackrel{\text{def}}{=} \arg \max_{\theta} \text{Lik}(\theta) \tag{4.2}$$

In nearly all cases, the MLE is consistent (Cramer, 1946), and gives intuitive results. In many common cases, it is also unbiased. For estimation of multinomial probabilities, the MLE also turns out to be the relative-frequency estimate. Figure 4.2 visualizes an example of this. The MLE is also an intuitive and unbiased estimator for the means of normal and Poisson distributions.

### 4.3.2 Limitations of the MLE: variance

As intuitive and general-purpose as it may be, the MLE has several important limitations, hence there is more to statistics than maximum-likelihood. Although the MLE for multinomial distributions is unbiased, its variance is problematic for estimating parameters that determine probabilities of events with low expected counts. This can be a major problem even when the sample size is very large. For example, WORD N-GRAM PROBABILITIES—the probability distribution over the next word in a text given the previous  $n - 1$  words of context—are of major interest today not only in applied settings such as speech recognition but also in the context of theoretical questions regarding language production, comprehension, and acquisition (e.g., Gahl, 2008; Saffran et al., 1996b; 2-gram probabilities are

---

<sup>2</sup>The expression  $\arg \max_x f(x)$  is defined as “the value of  $x$  that yields the maximum value for the expression  $f(x)$ .” It can be read as “arg-max over  $x$  of  $f(x)$ .”

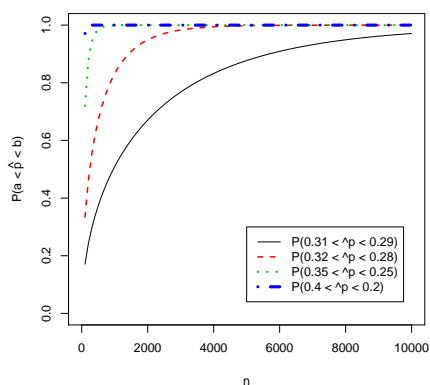


Figure 4.1: Consistency of relative frequency estimation. Plot indicates the probability with which the relative-frequency estimate  $\hat{\pi}$  for binomial distribution with parameter  $\pi = 0.3$  lies in narrow ranges around the true parameter value as a function of sample size  $n$ .

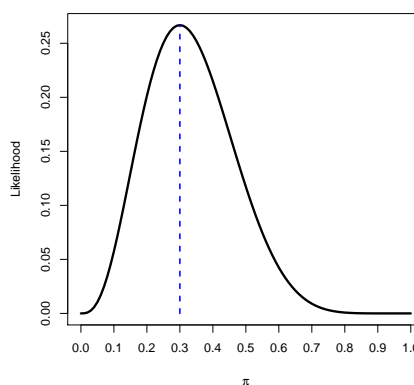


Figure 4.2: The likelihood function for the binomial parameter  $\pi$  for observed data where  $n = 10$  and  $m = 3$ . The MLE is the RFE for the binomial distribution. Note that this graph is *not* a probability density and the area under the curve is much less than 1.

sometimes called TRANSITIONAL PROBABILITIES). N-gram probability models are simply collections of large multinomial distributions (one distribution per context). Yet even for extremely high-frequency preceding contexts, such as the word sequence *near the*, there will be many possible next words that are improbable yet not impossible (for example, *reportedly*). Any word that does not appear in the observed data in that context will be assigned a conditional probability of zero by the MLE. In a typical n-gram model there will be many, many such words—the problem of DATA SPARSITY. This means that the MLE is a terrible means of prediction for n-gram word models, because if *any* unseen word continuation appears in a new dataset, the MLE will assign zero likelihood to the *entire dataset*. For this reason, there is a substantial literature on learning high-quality n-gram models, all of which can in a sense be viewed as managing the variance of estimators for these models while keeping the bias reasonably low (see Chen and Goodman, 1998 for a classic survey).

### 4.3.3 Limitations of the MLE: bias

In addition to these problems with variance, the MLE is biased for some types of model parameters. Imagine a linguist interested in inferring the original time of introduction of a novel linguistic expression currently in use today, such as the increasingly familiar phrase *the boss of me*, as in:<sup>3</sup>

- (3) “You’re too cheeky,” said Astor, sticking out his tongue. “You’re not the boss of me.”

<sup>3</sup>This phrase has been the topic of intermittent discussion on the *Language Log* blog since 2007.

(Tool, 1949, cited in *Language Log* by Benjamin Zimmer, 18 October 2007)

The only direct evidence for such expressions is, of course, attestations in written or recorded spoken language. Suppose that the linguist had collected 60 attestations of the expression, the oldest of which was recored 120 years ago.

From a probabilistic point of view, this problem involves choosing a probabilistic model whose generated observations are  $n$  attestation dates  $\mathbf{y}$  of the linguistic expression, and one of whose parameters is the earliest time at which the expression is coined, or  $t_0$ . When the problem is framed this way, the linguist's problem is to devise a procedure for constructing a parameter estimate  $\hat{t}_0$  from observations. For expository purposes, let us oversimplify and use the uniform distribution as a model of how attestation dates are generated.<sup>4</sup> Since the innovation is still in use today (time  $t_{now}$ ), the parameters of the uniform distribution are  $[t_0, t_{now}]$  and the only parameter that needs to be estimated is  $t_{now}$ . Let us arrange our attestation dates in chronological order so that the earliest date is  $y_1$ .

What is the maximum-likelihood estimate  $\hat{t}_0$ ? For a given choice of  $t_0$ , a given date  $y_i$  either falls in the interval  $[t_0, t_{now}]$  or it does not. From the definition of the uniform distribution (Section 2.7.1) we have:

$$P(y_i|t_0, t_{now}) = \begin{cases} \frac{1}{t_{now}-t_0} & t_0 \leq y_i \leq t_{now} \\ 0 & \text{otherwise} \end{cases} \quad (4.3)$$

Due to independence, the likelihood for the interval boundaries is  $\text{Lik}(t_0) = \prod_i P(y_i|t_0, t_{now})$ . This means that for any choice of interval boundaries, if at least one date lies before  $t_0$ , the entire likelihood is zero! Hence the likelihood is non-zero only for interval boundaries containing all dates. For such boundaries, the likelihood is

$$\text{Lik}(t_0) = \prod_{i=1}^n \frac{1}{t_{now} - t_0} \quad (4.4)$$

$$= \frac{1}{(t_{now} - t_0)^n} \quad (4.5)$$

This likelihood grows larger as  $t_{now} - t_0$  grows smaller, so it will be maximized when the interval length  $t_{now} - t_0$  is as short as possible—namely, when  $t_0$  is set to the earliest attested date  $y_1$ . This fact is illustrated in Figure 4.3: the tighter the posited interval between  $t_0$  and  $t_{now}$ , the greater the resulting likelihood.

You probably have the intuition that this estimate of the contact interval duration is conservative: certainly the novel form appeared in English no later than  $t_0$ , but it seems

---

<sup>4</sup>This is a dramatic oversimplification, as it is well known that linguistic innovations prominent enough for us to notice today often followed an S-shaped trajectory of usage frequency (Bailey, 1973; Cavall-Sforza and Feldman, 1981; Kroch, 1989; Wang and Minnett, 2005). However, the general issue of bias in maximum-likelihood estimation present in the oversimplified uniform-distribution model here also carries over to more complex models of the diffusion of linguistic innovations.

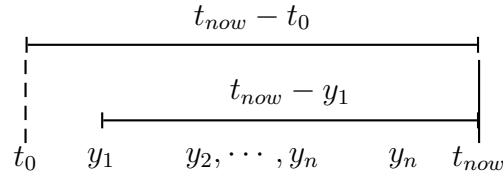


Figure 4.3: The bias of the MLE for uniform distributions

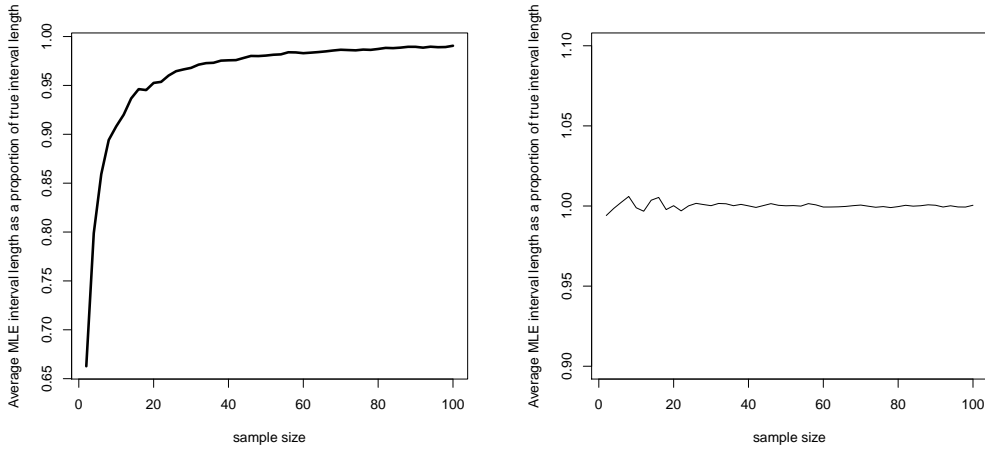


Figure 4.4: Bias of the MLE (left) and the bias-corrected estimator (right), shown numerically using 500 simulations for each value of  $n$ .

rather unlikely that the first use in the language was also the first attested use!<sup>5</sup> This intuition is correct, and its mathematical realization is that the MLE for interval boundaries of a uniform distribution is biased. Figure 4.4 visualizes this bias in terms of average interval length (over a number of samples) as a function of sample size.

For any finite sample size, the MLE is biased to underestimate true interval length, although this bias decreases as sample size increases (as well it should, because the MLE is a consistent estimator). Fortunately, the size of the MLE's bias can be quantified analytically: the expected ML-estimated interval size is  $\frac{n}{n+1}$  times the true interval size. Therefore, if we adjust the MLE by multiplying it by  $\frac{n+1}{n}$ , we arrive at an unbiased estimator for interval length. The correctness of this adjustment is confirmed by the right-hand plot in Figure 4.4. In the case of our historical linguist with three recovered documents, we achieve the estimate

$$\hat{t}_0 = 120 \times \frac{61}{60} = 122 \text{ years ago}$$

<sup>5</sup>The intuition may be different if the first attested use was by an author who is known to have introduced a large number of novel expressions into the language which subsequently gained in popularity. This type of situation would point to a need for a more sophisticated probabilistic model of innovation, diffusion, and attestation.

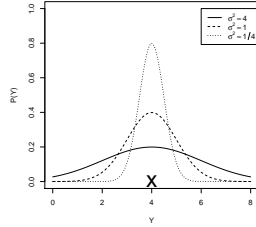


Figure 4.5: Bias in the MLE for  $\sigma$  of a normal distribution.

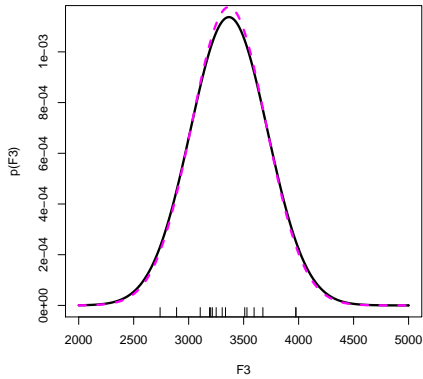


Figure 4.6: Point estimation of a normal distribution. The maximum-likelihood estimate is the dotted magenta line; the bias-adjusted estimate is solid black.

Furthermore, there is a degree of intuitiveness about the behavior of the adjustment in extreme cases: if  $N = 1$ , the adjustment would be infinite, which makes sense: one cannot estimate the size of an unconstrained interval from a single observation.

Another famous example of bias in the MLE is in estimating the variance of a normal distribution. The MLEs for mean and variance of a normal distribution as estimated from a set of  $N$  observations  $\mathbf{y}$  are as follows:

$$\hat{\mu}_{MLE} = \frac{1}{N} \sum_i y_i \quad (\text{i.e. the sample mean}) \quad (4.6)$$

$$\hat{\sigma}_{MLE}^2 = \frac{1}{N} \sum_i (y_i - \hat{\mu})^2 \quad (\text{i.e. the sample variance divided by } N) \quad (4.7)$$

While it turns out that  $\hat{\mu}_{MLE}$  is unbiased,  $\hat{\sigma}_{MLE}^2$  is biased for reasons similar to those given for interval size in the uniform distribution. You can see this graphically by imagining the MLE for a single observation, as in Figure 4.5. As  $\hat{\sigma}^2$  shrinks, the likelihood of the observation will continue to rise, so that the MLE will push the estimated variance to be arbitrarily small. This is a type of **OVERFITTING** (see Section 2.11.5).

It turns out that this bias can be eliminated by adjusting the MLE by the factor  $\frac{N}{N-1}$ . This adjusted estimate of  $\sigma^2$  is called  $S^2$ :

$$S^2 = \frac{N}{N-1} \hat{\sigma}_{MLE}^2 \quad (4.8)$$

$$= \frac{1}{N-1} \sum_i (y_i - \hat{\mu})^2 \quad (4.9)$$

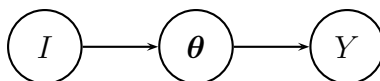


Figure 4.7: The structure of a simple Bayesian model. Observable data  $Y$  and prior beliefs  $I$  are conditionally independent given the model parameters.

This is the most frequently used estimate of the underlying variance of a normal distribution from a sample. In R, for example, the function `var()`, which is used to obtain sample variance, computes  $S^2$  rather than  $\hat{\sigma}_{MLE}$ . An example of estimating normal densities is shown in Figure 4.6, using F3 formants from 15 native English-speaking children on the vowel [æ]. The MLE density estimate is a slightly narrower curve than the bias-adjusted estimate.

## 4.4 Bayesian parameter estimation and density estimation

In frequentist statistics as we have discussed thus far, one uses observed data to construct a point estimate for each model parameter. The MLE and bias-adjusted version of the MLE are examples of this. In Bayesian statistics, on the other hand, parameter estimation involves placing a *probability distribution* over model parameters. In fact, there is no conceptual difference between parameter estimation (inferences about  $\theta$ ) and prediction or density estimation (inferences about future  $\mathbf{y}$ ) in Bayesian statistics.

### 4.4.1 Anatomy of inference in a simple Bayesian model

A simple Bayesian model has three components. Observable data are generated as random variables  $\mathbf{y}$  in some model from a model family with parameters  $\theta$ . Prior to observing a particular set of data, however, we already have beliefs/expectations about the possible model parameters  $\theta$ ; we call these beliefs  $I$ . These beliefs affect  $\mathbf{y}$  only through the mediation of the model parameters—that is,  $\mathbf{y}$  and  $I$  are *conditionally independent* given  $\theta$  (see Section 2.4.2). This situation is illustrated in Figure 6.1, which has a formal interpretation as a graphical model (Appendix C).

In the Bayesian framework, both parameter estimation and density estimation simply involve the application of Bayes’ rule (Equation (2.5)). For example, parameter estimation means calculating the probability distribution over  $\theta$  given observed data  $\mathbf{y}$  and our prior beliefs  $I$ . We can use Bayes rule to write this distribution as follows:



$$P(\boldsymbol{\theta}|\mathbf{y}, I) = \frac{P(\mathbf{y}|\boldsymbol{\theta}, I)P(\boldsymbol{\theta}|I)}{P(\mathbf{y}|I)} \quad (4.10)$$

$$= \frac{\underbrace{P(\mathbf{y}|\boldsymbol{\theta})}_{\text{Likelihood for } \boldsymbol{\theta}} \underbrace{P(\boldsymbol{\theta}|I)}_{\text{Prior over } \boldsymbol{\theta}}}{\underbrace{P(\mathbf{y}|I)}_{\text{Likelihood marginalized over } \boldsymbol{\theta}}} \quad (\text{because } \mathbf{y} \perp I \mid \boldsymbol{\theta}) \quad (4.11)$$

The numerator in Equation (4.11) is composed of two quantities. The first term,  $P(\mathbf{y}|\boldsymbol{\theta})$ , should be familiar from Section 2.11.5: it is the likelihood of the parameters  $\boldsymbol{\theta}$  for the data  $\mathbf{y}$ . As in much of frequentist statistics, the likelihood plays a central role in parameter estimation in Bayesian statistics. However, there is also a second term,  $P(\boldsymbol{\theta}|I)$ , the **PRIOR DISTRIBUTION** over  $\boldsymbol{\theta}$  given only  $I$ . The complete quantity (4.11) is the **POSTERIOR DISTRIBUTION** over  $\boldsymbol{\theta}$ . It is important to realize that the terms “prior” and “posterior” in no way imply any temporal ordering on the realization of different events. The only thing that  $P(\boldsymbol{\theta}|I)$  is “prior” to is the incorporation of the particular dataset  $\mathbf{y}$  into inferences about  $\boldsymbol{\theta}$ .  $I$  can in principle incorporate all sorts of knowledge, including other data sources, scientific intuitions, or—in the context of language acquisition—innate biases. Finally, the denominator is simply the **MARGINAL LIKELIHOOD**  $P(\mathbf{y}|I) = \int_{\boldsymbol{\theta}} P(\mathbf{y}|\boldsymbol{\theta})P(\boldsymbol{\theta}|I) d\boldsymbol{\theta}$  (it is the model parameters  $\boldsymbol{\theta}$  that are being marginalized over; see Section 3.2). The data likelihood is often the most difficult term to calculate, but in many cases its calculation can be ignored or circumvented because we can accomplish everything we need by computing posterior distributions up to a normalizing constant (Section 2.8; we will see an new example of this in the next section).

Since Bayesian inference involves placing probability distributions on model parameters, it becomes useful to work with probability distributions that are specialized for this purpose. Before we move on to our first simple example of Bayesian parameter and density estimation, we’ll now introduce one of the simplest (and most easily interpretable) such probability distributions: the beta distribution.

#### 4.4.2 The beta distribution

The **BETA DISTRIBUTION** is important in Bayesian statistics involving binomial distributions. It has two parameters  $\alpha_1, \alpha_2$  and is defined as follows:

$$P(\pi|\alpha_1, \alpha_2) = \frac{1}{B(\alpha_1, \alpha_2)} \pi^{\alpha_1-1} (1-\pi)^{\alpha_2-1} \quad (0 \leq \pi \leq 1, \alpha_1 > 0, \alpha_2 > 0) \quad (4.12)$$

where the **BETA FUNCTION**  $B(\alpha_1, \alpha_2)$  (Section B.1) serves as a normalizing constant:

$$B(\alpha_1, \alpha_2) = \int_0^1 \pi^{\alpha_1-1} (1-\pi)^{\alpha_2-1} d\pi \quad (4.13)$$

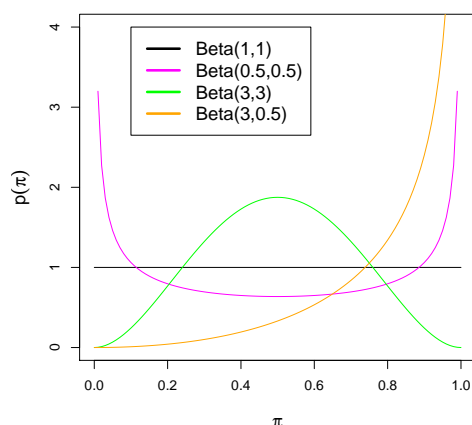


Figure 4.8: Beta distributions

Figure 4.8 gives a few examples of beta densities for different parameter choices. The beta distribution has a mean of  $\frac{\alpha_1}{\alpha_1 + \alpha_2}$  and mode (when both  $\alpha_1, \alpha_2 > 1$ ) of  $\frac{\alpha_1 - 2}{\alpha_1 + \alpha_2 - 2}$ . Note that a uniform distribution on  $[0, 1]$  results when  $\alpha_1 = \alpha_2 = 1$ .

Beta distributions and beta functions are very often useful when dealing with Bayesian inference on binomially-distributed data. One often finds oneself in the situation of knowing that some random variable  $X$  is distributed such that  $P(X) \propto \pi^a(1-\pi)^b$ , but not knowing the normalization constant. If and when you find yourself in this situation, recognize that  $X$  must be beta-distributed, which allows you to determine the normalization constant immediately. Additionally, whenever one is confronted with an integral of the form  $\int_0^1 \pi^a(1-\pi)^b d\pi$  (as in Section 5.2.1), recognize that it is a beta function, which will allow you to compute the integral very easily.

### 4.4.3 Simple example of Bayesian estimation with the binomial distribution

Historically, one of the major reasons that Bayesian inference has been avoided is that it can be computationally intensive under many circumstances. The rapid improvements in available computing power over the past few decades are, however, helping overcome this obstacle, and Bayesian techniques are becoming more widespread both in practical statistical applications and in theoretical approaches to modeling human cognition. We will see examples of more computationally intensive techniques later in the book, but to give the flavor of the Bayesian approach, let us revisit the example of our native American English speaker and her quest for an estimator for  $\pi$ , the probability of the passive voice, which turns out to be analyzable without much computation at all.

We have already established that transitive sentences in the new variety can be modeled using a binomial distribution where the parameter  $\pi$  characterizes the probability that a

given transitive sentence will be in the passive voice. For Bayesian statistics, we must first specify the beliefs  $I$  that characterize the prior distribution  $P(\boldsymbol{\theta}|I)$  to be held before any data from the new English variety is incorporated. In principle, we could use any proper probability distribution on the interval  $[0, 1]$  for this purpose, but here we will use the beta distribution (Section 4.4.2). In our case, specifying prior knowledge  $I$  amounts to choosing beta distribution parameters  $\alpha_1$  and  $\alpha_2$ .

Once we have determined the prior distribution, we are in a position to use a set of observations  $\mathbf{y}$  to do parameter estimation. Suppose that the observations  $\mathbf{y}$  that our speaker has observed are comprised of  $n$  total transitive sentences,  $m$  of which are passivized. Let us simply instantiate Equation (4.11) for our particular problem:

$$P(\pi|\mathbf{y}, \alpha_1, \alpha_2) = \frac{P(\mathbf{y}|\pi)P(\pi|\alpha_1, \alpha_2)}{P(\mathbf{y}|\alpha_1, \alpha_2)} \quad (4.14)$$

The first thing to notice here is that the denominator,  $P(\mathbf{y}|\alpha_1, \alpha_2)$ , is not a function of  $\pi$ . That means that it is a normalizing constant (Section 2.8). As noted in Section 4.4, we can often do everything we need without computing the normalizing constant, here we ignore the denominator by re-expressing Equation (4.14) in terms of proportionality:

$$P(\pi|\mathbf{y}, \alpha_1, \alpha_2) \propto P(\mathbf{y}|\pi)P(\pi|\alpha_1, \alpha_2)$$

From what we know about the binomial distribution, the likelihood is  $P(\mathbf{y}|\pi) = \binom{n}{m}\pi^m(1 - \pi)^{n-m}$ , and from what we know about the beta distribution, the prior is  $P(\pi|\alpha_1, \alpha_2) = \frac{1}{B(\alpha_1, \alpha_2)}\pi^{\alpha_1-1}(1 - \pi)^{\alpha_2-1}$ . Neither  $\binom{n}{m}$  nor  $B(\alpha_1, \alpha_2)$  is a function of  $\pi$ , so we can also ignore them, giving us

$$\begin{aligned} P(\pi|\mathbf{y}, \alpha_1, \alpha_2) &\propto \overbrace{\pi^m(1 - \pi)^{n-m}}^{\text{Likelihood}} \overbrace{\pi^{\alpha_1-1}(1 - \pi)^{\alpha_2-1}}^{\text{Prior}} \\ &\propto \pi^{m+\alpha_1-1}(1 - \pi)^{n-m+\alpha_2-1} \end{aligned} \quad (4.15)$$

Now we can crucially notice that the posterior distribution on  $\pi$  itself has the form of a beta distribution (Equation (4.12)), with parameters  $\alpha_1 + m$  and  $\alpha_2 + n - m$ . This fact that the posterior has the same functional form as the prior is called CONJUGACY; the beta distribution is said to be CONJUGATE TO the binomial distribution. Due to conjugacy, we can circumvent the work of directly calculating the normalizing constant for Equation (4.15), and recover it from what we know about beta distributions. This gives us a normalizing constant of  $B(\alpha_1 + m, \alpha_2 + n - m)$ .

Now let us see how our American English speaker might apply Bayesian inference to estimating the probability of passivization in the new English variety. A reasonable prior distribution might involve assuming that the new variety could be somewhat like American English. Approximately 8% of spoken American English sentences with simple transitive

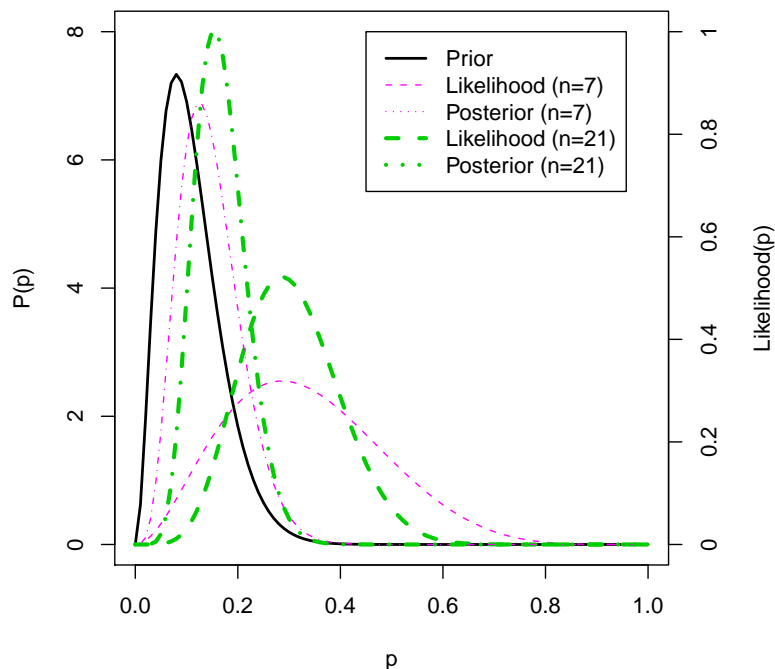


Figure 4.9: Prior, likelihood, and posterior distributions over  $\pi$ . Note that the likelihood has been rescaled to the scale of the prior and posterior; the original scale of the likelihood is shown on the axis on the right.

verbs are passives (Roland et al., 2007), hence our speaker might choose  $\alpha_1$  and  $\alpha_2$  such that the mode of  $P(\pi|\alpha_1, \alpha_2)$  is near 0.08. A beta distribution has a mode if  $\alpha_1, \alpha_2 > 1$ , in which case the mode is  $\frac{\alpha_1-1}{\alpha_1+\alpha_2-2}$ , so a reasonable choice might be  $\alpha_1 = 3, \alpha_2 = 24$ , which puts the mode of the prior distribution at  $\frac{2}{25} = 0.08$ .<sup>6</sup> Now suppose that our speaker is exposed to  $n = 7$  transitive verbs in the new variety, and two are passivized ( $m = 2$ ). The posterior distribution will then be beta-distributed with  $\alpha_1 = 3 + 2 = 5, \alpha_2 = 24 + 5 = 29$ . Figure 4.9 shows the prior distribution, likelihood, and posterior distribution for this case, and also for the case where the speaker has been exposed to three times as much data in similar proportions ( $n = 21, m = 6$ ). In the  $n = 7$ , because the speaker has seen relatively little data, the prior distribution is considerably more peaked than the likelihood, and the posterior distribution is fairly close to the prior. However, as our speaker sees more and more data, the likelihood becomes increasingly peaked, and will eventually dominate in the behavior of the posterior (See Exercise to be included with this chapter).

In many cases it is useful to summarize the posterior distribution into a point estimate of the model parameters. Two commonly used such point estimates are the *mode* (which

<sup>6</sup>Compare with Section 4.3.1—the binomial likelihood function has the same shape as a beta distribution!

we covered a moment ago) and the *mean*. For our example, the posterior mode is  $\frac{4}{32}$ , or 0.125. Selecting the mode of the posterior distribution goes by the name of MAXIMUM A POSTERIORI (MAP) estimation. The mean of a beta distribution is  $\frac{\alpha_1}{\alpha_1 + \alpha_2}$ , so our POSTERIOR MEAN is  $\frac{5}{34}$ , or about 0.15. There are no particular deep mathematical principles motivating the superiority of the mode over the mean or vice versa, although the mean should generally be avoided in cases where the posterior distribution is multimodal. The most “principled” approach to Bayesian parameter estimation is in fact *not* to choose a point estimate for model parameters after observing data, but rather to make use of the entire posterior distribution in further statistical inference.

## Bayesian density estimation

The role played in density estimation by parameter estimation up to this point has been as follows: an estimator is applied to observed data to obtain an estimate for model parameters  $\hat{\theta}$ , and the resulting probabilistic model determines a set of predictions for future data, namely the distribution  $P(Y|\hat{\theta})$ . If we use Bayesian inference to form a posterior distribution on  $\theta$  and then summarize that distribution into a point estimate, we can use that point estimate in exactly the same way. In this sense, using a given prior distribution together with the MAP or posterior mean can be thought of as simply one more estimator. In fact, this view creates a deep connection between Bayesian inference and maximum-likelihood estimation: maximum-likelihood estimation (Equation (4.2)) is simply Bayesian MAP estimation when the prior distribution  $P(\theta|I)$  (Equation (4.10)) is taken to be uniform over all values of  $\theta$ .

However, in the purest Bayesian view, it is undesirable to summarize our beliefs about model parameters into a point estimate, because this discards information. In Figure 4.9, for example, the two likelihoods are peaked at the same place, but the  $n = 21$  likelihood is more peaked than the  $n = 7$  likelihood. This translates into more peakedness and therefore more certainty in the posterior; this certainty is not reflected in the MLE or even in the MAP estimate. Pure Bayesian density estimation involves *marginalization* (Section 3.2) over the model parameters, a process which automatically incorporates this degree of certainty. That is, we estimate a density over new observations  $\mathbf{y}_{new}$  as:

$$P(\mathbf{y}_{new}|\mathbf{y}, I) = \int_{\theta} P(\mathbf{y}_{new}|\theta)P(\theta|\mathbf{y}, I) d\theta \quad (4.16)$$

where  $P(\theta|\mathbf{y}, I)$  is familiar from Equation (4.11).

Suppose, for example, that after hearing her  $n$  examples from the new English dialect, our speaker wanted to predict the number of passives  $r$  she would hear after the next  $k$  trials. We would have:

$$P(r|k, I, \mathbf{y}) = \int_0^1 P(r|k, \pi)P(\pi|\mathbf{y}, I) d\pi$$

This expression can be reduced to

$$P(r|k, I, \mathbf{y}) = \binom{k}{r} \frac{\prod_{i=0}^{r-1} (\alpha_1 + m + i) \prod_{i=0}^{k-r-1} (\alpha_2 + n - m + i)}{\prod_{i=0}^{k-1} (\alpha_1 + \alpha_2 + n + i)}$$

which is known as the BETA-BINOMIAL MODEL. The expression may seem formidable, but experimenting with specific values for  $k$  and  $r$  reveals that it is simpler than it may seem. For a single trial ( $k = 1$ ), for example, this expression reduces to  $P(r = 1|k, I, \mathbf{y}) = \frac{\alpha_1 + m}{\alpha_1 + \alpha_2 + n}$ , which is exactly what would be obtained by using the posterior mean. For two trials ( $k = 2$ ), we would have  $P(r = 1|k, I, \mathbf{y}) = \frac{(\alpha_1 + m)(\alpha_2 + n - m)}{(\alpha_1 + \alpha_2 + n)(\alpha_1 + \alpha_2 + n + 1)}$ , which is slightly less than what would be obtained by using the posterior mean.<sup>7</sup> This probability mass lost from the  $r = 1$  outcome is redistributed into the more extreme  $r = 0$  and  $r = 2$  outcomes. For  $k > 1$  trials in general, the beta-binomial model leads to density estimates of greater variance—also called DISPERSION in the modeling context—than for the binomial model using posterior mean. This is illustrated in Figure 4.10. The reason for this greater dispersion is that different future trials are only conditionally independent given a fixed choice of the binomial parameter  $\pi$ . Because there is residual uncertainty about this parameter, successes on different future trials are positively correlated in the Bayesian prediction despite the fact that they are conditionally independent given the underlying model parameter (see also Section 2.4.2 and Exercise 2.2). This is an important property of a wide variety of models which involve marginalization over intermediate variables (in this case the binomial parameter); we will return to this in Chapter 8 and later in the book.

## 4.5 Computing approximate Bayesian inferences with sampling techniques

In the example of Bayesian inference given in Section 4.4.3, we were able to express both (i) the posterior probability over the binomial parameter  $\pi$ , and (ii) the probability distribution over new observations as the CLOSED-FORM expressions<sup>8</sup> shown in Equations (4.15) and (4.17) respectively. We were able to do this due to the CONJUGACY of the beta distribution to the binomial distribution. However, it will sometimes be the case that we want to perform Bayesian inferences but don't have conjugate distributions to work with. As a simple example, let us turn back to a case of inferring the ordering preference of an English

---

<sup>7</sup>With the posterior mean, the term  $(\alpha_1 + \alpha_2 + n + 1)$  in the denominator would be replaced by another instance of  $(\alpha_1 + \alpha_2 + n)$ , giving us

$$P(r = 1|k, \hat{\pi}) = \frac{(\alpha_1 + m)(\alpha_2 + n - m)}{(\alpha_1 + \alpha_2 + n)^2} \quad (4.17)$$

<sup>8</sup>A closed-form expression is one that can be written exactly as a combination of a finite number of “well-known” functions (such as polynomials, logarithms, exponentials, and so forth).

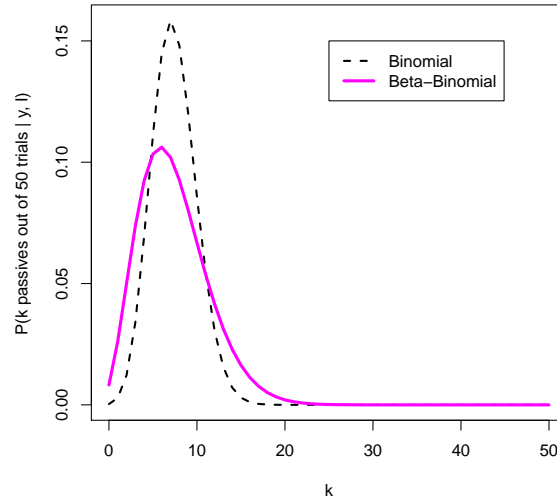


Figure 4.10: The beta-binomial model has greater dispersion than the binomial model. Results shown for  $\alpha_1 + m = 5$ ,  $\alpha_2 + n - m = 29$ .

binomial, such as  $\{\textit{radio}, \textit{television}\}$ . The words in this particular binomial differ in length (quantified as, for example, number of syllables), and numerous authors have suggested that a short-before-long metrical constraint is one determinant of ordering preferences for English binomials (Cooper and Ross, 1975; Pinker and Birdsong, 1979, *inter alia*). Our prior knowledge therefore inclines us to expect a preference for the ordering *radio and television* (abbreviated as  $\mathbf{r}$ ) more strongly than a preference for the ordering *television and radio* ( $\mathbf{t}$ ), but we may be relatively agnostic as to the particular strength of the ordering preference. A natural probabilistic model here would be the binomial distribution with success parameter  $\pi$ , and a natural prior might be one which is uniform within each of the ranges  $0 \leq \pi \leq 0.5$  and  $0.5 < \pi < 1$ , but twice as large in the latter range as in the former range. This would be the following prior:

$$p(\pi = x) = \begin{cases} \frac{2}{3} & 0 \leq x \leq 0.5 \\ \frac{4}{3} & 0.5 < x \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad (4.18)$$

which is a step function, illustrated in Figure 4.11a.

In such cases, there are typically no closed-form expressions for the posterior or predictive distributions given arbitrary observed data  $\mathbf{y}$ . However, these distributions can very often be approximated using general-purpose SAMPLING-BASED approaches. Under these approaches, samples (in principle independent of one another) can be drawn over quantities that are unknown in the model. These samples can then be used in combination with density

estimation techniques such as those from Chapter ?? to approximate any probability density of interest. Chapter ?? provides a brief theoretical and practical introduction to sampling techniques; here, we introduce the steps involved in sampling-based approaches as needed.

For example, suppose we obtain data  $\mathbf{y}$  consisting of ten binomial tokens—five of  $\mathbf{r}$  and five of  $\mathbf{t}$ —and are interested in approximating the following distributions:

1. The posterior distribution over the success parameter  $\pi$ ;
2. The posterior predictive distribution over the observed ordering of an eleventh token;
3. The posterior predictive distribution over the number of  $\mathbf{r}$  orderings seen in ten more tokens.

We can use BUGS, a highly flexible language for describing and sampling from structured probabilistic models, to sample from these distributions. BUGS uses GIBBS SAMPLING, a Markov-chain Monte Carlo technique (Chapter ??), to produce samples from the posterior distributions of interest to us (such as  $P(\pi|\mathbf{y}, I)$  or  $P(\mathbf{y}_{new}|\mathbf{y}, I)$ ). Here is one way to describe our model in BUGS:

```
model {
  /* the model */
  for(i in 1:length(response)) { response[i] ~ dbern(p) }
  /* the prior */
  pA ~ dunif(0,0.5)
  pB ~ dunif(0.5,1)
  i ~ dbern(2/3)
  p <- (1 - i) * pA + i * pB
  /* predictions */
  prediction1 ~ dbern(p)
  prediction2 ~ dbin(p, 10) /* dbin() is for binomial distribution */
}
```

The first line,

```
for(i in 1:length(response)) { response[i] ~ dbern(p) }
```

says that each observation is the outcome of a Bernoulli random variable with success parameter  $p$ .

The next part,

```
pA ~ dunif(0,0.5)
pB ~ dunif(0.5,1)
i ~ dbern(2/3)
p <- (1 - i) * pA + i * pB
```



is a way of encoding the step-function prior of Equation (4.18). The first two lines say that there are two random variables,  $p_A$  and  $p_B$ , drawn from uniform distributions on  $[0, 0.5]$  and  $[0.5, 1]$  respectively. The next two lines say that the success parameter  $p$  is equal to  $p_A \frac{2}{3}$  of the time, and is equal to  $p_B$  otherwise. These four lines together encode the prior of Equation (4.18).

Finally, the last two lines say that there are two more random variables parameterized by  $p$ : a single token (`prediction1`) and the number of  $r$  outcomes in ten more tokens (`prediction2`).

There are several incarnations of BUGS, but here we focus on a newer incarnation, JAGS, that is open-source and cross-platform. JAGS can interface with R through the R library `rjags`.<sup>9</sup> Below is a demonstration of how we can use BUGS through R to estimate the posteriors above with samples.

```
> rm(list=ls())
> set.seed(45)
> library(rjags)

module basemod loaded
module bugs loaded

> # first, set up observed data
> response <- c(rep(1,5),rep(0,5))
> # now compile the BUGS model
> m <- jags.model("../jags_examples/asymm_binomial_prior/asymm_binomial_prior.bug")

Compiling model graph
  Resolving undeclared variables
  Allocating nodes
  Graph Size: 26

> # initial period of running the model to get it converged
> update(m,1000)
> # Now get samples
> res <- coda.samples(m, c("p","prediction1","prediction2"), thin = 20, n.iter=5000)
> # posterior predictions not completely consistent due to sampling noise
> print(apply(res[[1]],2,mean))

      p prediction1 prediction2
0.5449781  0.5000000  5.5360000

> posterior.mean <- apply(res[[1]],2,mean)
```

---

<sup>9</sup>JAGS can be obtained freely at <http://calvin.iarc.fr/~martyn/software/jags/>, and `rjags` at <http://cran.r-project.org/web/packages/rjags/index.html>.

```

> plot(density(res[[1]][, 1]), xlab = expression(pi), ylab = expression(paste("p(",
+   pi, ")")))

> preds2 <- table(res[[1]][, 3])
> plot(preds2/sum(preds2), type = "h", xlab = "r", ylab = "P(r|y)",
+   lwd = 4, ylim = c(0, 0.25))
> posterior.mean.predicted.freqs <- dbinom(0:10, 10, posterior.mean[1])
> x <- 0:10 + 0.1
> arrows(x, 0, x, posterior.mean.predicted.freqs, length = 0, lty = 2,
+   lwd = 4, col = "magenta")
> legend(0, 0.25, c(expression(paste("Marginizing over ", pi)),
+   "With posterior mean"), lty = c(1, 2), col = c("black", "magenta"),
+   lwd = 4)

```

Two important notes on the use of sampling: first, immediately after compiling we specify a “burn-in” period of 1000 iterations to bring the Markov chain to a “steady state” with:<sup>10</sup>

```
update(m, 1000)
```

Second, there can be AUTOCORRELATION in the Markov chain: samples near to one another in time are non-independent of one another.<sup>11</sup> In order to minimize the bias in the estimated probability density, we’d like to minimize this autocorrelation. We can do this by sub-sampling or “thinning” the Markov chain, in this case taking only one out of every 20 samples from the chain as specified by the argument `thin = 20` to `coda.samples()`. This reduces the autocorrelation to a minimal level. We can get a sense of how bad the autocorrelation is by taking an unthinned sample and computing the autocorrelation at a number of time lags:

```

> res <- coda.samples(m, c("p", "prediction1", "prediction2"),
+   thin = 1, n.iter = 5000)
> autocorr(res, lags = c(1, 5, 10, 20, 50))

```

```
[[1]]
, , p
```

|        | p           | prediction1  | prediction2 |
|--------|-------------|--------------|-------------|
| Lag 1  | 0.22281273  | 0.0481639827 | 0.14894624  |
| Lag 5  | 0.01597045  | 0.0008096763 | 0.01299937  |
| Lag 10 | 0.02254423  | 0.0078164724 | 0.01473938  |
| Lag 20 | -0.02553418 | 0.0190368737 | -0.03753145 |

<sup>10</sup>For any given model there is no guarantee how many iterations are needed, but most of the models covered in this book are simple enough that on the order of thousands of iterations is enough.

<sup>11</sup>The autocorrelation of a sequence  $\vec{x}$  for a time lag  $\tau$  is simply the covariance between elements in the sequence that are  $\tau$  steps apart, or  $\text{Cov}(x_i, x_{i+\tau})$ .

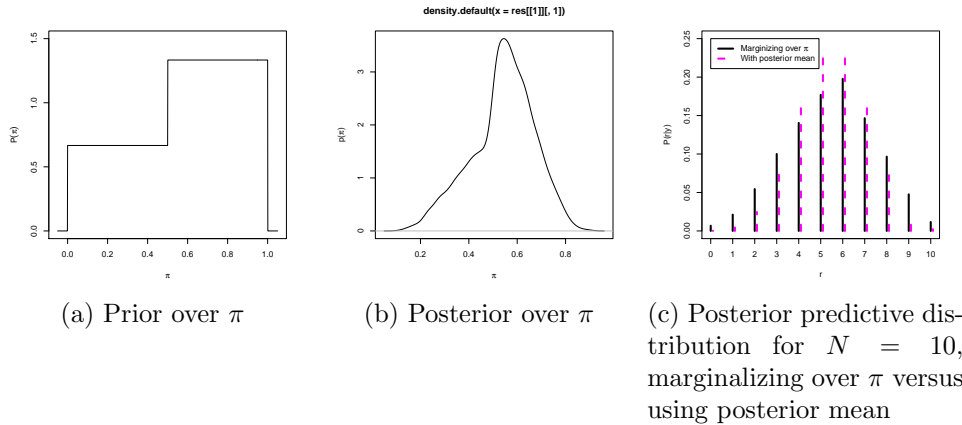


Figure 4.11: A non-conjugate prior for the binomial distribution: prior distribution, posterior over  $\pi$ , and predictive distribution for next 10 outcomes

```
Lag 50 -0.01487515 -0.0096603093 -0.01116133
```

```
, , prediction1
```

|        | p           | prediction1  | prediction2 |
|--------|-------------|--------------|-------------|
| Lag 1  | 0.05018839  | 0.013907531  | 0.04105389  |
| Lag 5  | 0.02350085  | 0.016018720  | 0.01962672  |
| Lag 10 | 0.02145659  | -0.001009159 | 0.00950255  |
| Lag 20 | -0.03309459 | -0.012707665 | -0.02598715 |
| Lag 50 | 0.01330366  | -0.007144728 | 0.01638643  |

```
, , prediction2
```

|        | p            | prediction1 | prediction2  |
|--------|--------------|-------------|--------------|
| Lag 1  | 0.138281803  | 0.047603265 | 0.073588715  |
| Lag 5  | 0.016129080  | 0.005837364 | 0.006626674  |
| Lag 10 | 0.016317285  | 0.017594950 | 0.013187233  |
| Lag 20 | -0.003506474 | 0.024783796 | -0.018888957 |
| Lag 50 | -0.010888272 | 0.003813316 | -0.005141463 |

We see that the autocorrelation is quite problematic for an unthinned chain (lag 1), but it is much better at higher lags. Thinning the chain by taking every twentieth sample is more than sufficient to bring the autocorrelation down

Notably, the posterior distribution shown in Figure 4.11a looks quite different from a beta distribution. Once again the greater dispersion of Bayesian prediction marginalizing over  $\pi$ , as compared with the predictions derived from the posterior mean, is evident in Figure 4.11c.

Finally, we'll illustrate one more example of simple Bayesian estimation, this time of a normal distribution for the F3 formant of the vowel [æ], based on speaker means of 15 child

native speakers of English from Peterson and Barney (1952). Since the normal distribution has two parameters—the mean  $\mu$  and variance  $\sigma^2$ —we must use a slightly more complex prior of the form  $P(\mu, \sigma^2)$ . We will assume that these parameters are independent of one another in the prior—that is,  $P(\mu, \sigma^2) = P(\mu)P(\sigma^2)$ . For our prior, we choose NON-INFORMATIVE distributions (ones that give similar probability to broad ranges of the model parameters). In particular, we choose uniform distributions over  $\mu$  and  $\log \sigma$  over the ranges  $[0, 10^5]$  and  $[-100, 100]$  respectively.<sup>12</sup> This gives us the model:

$$\begin{aligned} \mathbf{y} &\sim \mathcal{N}(\mu, \sigma^2) \\ \mu &\sim \mathcal{U}(0, 10^5) \\ \log \sigma &\sim \mathcal{U}(-100, 100) \end{aligned}$$

where  $\sim$  means “is distributed as”.

Here is the model in BUGS:

```
var predictions[M]
model {
  /* the model */
  for(i in 1:length(response)) { response[i] ~ dnorm(mu,tau) }
  /* the prior */
  mu ~ dunif(0,100000) # based on F3 means for other vowels
  log.sigma ~ dunif(-100,100)
  sigma <- exp(log.sigma)
  tau <- 1/(sigma^2)
  /* predictions */
  for(i in 1:M) { predictions[i] ~ dnorm(mu,tau) }
}
```

The first line,

```
var predictions[M]
```

states that the `predictions` variable will be a numeric array of length `M` (with `M` to be specified from `R`). BUGS parameterizes the normal distribution differently than we have, using a precision parameter  $\tau \stackrel{\text{def}}{=} \frac{1}{\sigma^2}$ . The next line,

```
for(i in 1:length(response)) { response[i] ~ dnorm(mu,tau) }
```

---

<sup>12</sup>See Gelman et al. (2004, Appendix C) for the relative merits of different choices of how to place a prior on  $\sigma^2$ .

simply expresses that observations  $\mathbf{y}$  are drawn from a normal distribution parameterized by  $\mu$  and  $\tau$ . The mean  $\mu$  is straightforwardly parameterized with a uniform distribution over a wide range. When we set the prior over  $\tau$  we do so in three stages, first saying that  $\log \sigma$  is uniformly distributed:

```
log.sigma ~ dunif(-100,100)
```

and transforming from  $\log \sigma$  to  $\sigma$  and then to  $\tau$ :

```
sigma <- exp(log.sigma)
tau <- 1/(sigma^2)
```

From R, we can compile the model and draw samples as before:

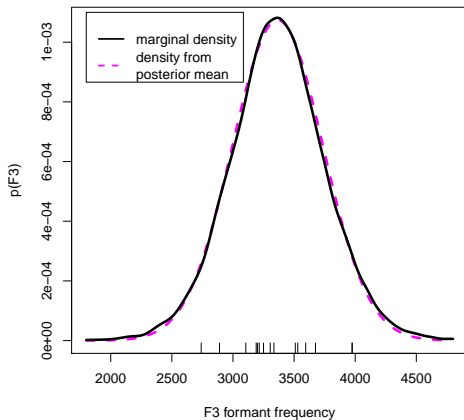
```
> pb <- read.table("../data/peterson_barney_data/pb.txt", header = T)
> pb.means <- with(pb, aggregate(data.frame(F0, F1, F2, F3), by = list(Type,
+   Sex, Speaker, Vowel, IPA), mean))
> names(pb.means) <- c("Type", "Sex", "Speaker", "Vowel", "IPA",
+   names(pb.means)[6:9])
> set.seed(18)
> response <- subset(pb.means, Vowel == "ae" & Type == "c")["F3"]
> M <- 10
> m <- jags.model("../jags_examples/child_f3_formant/child_f3_formant.bug")
```

```
Compiling model graph
  Declaring variables
  Resolving undeclared variables
  Allocating nodes
  Graph Size: 37
```

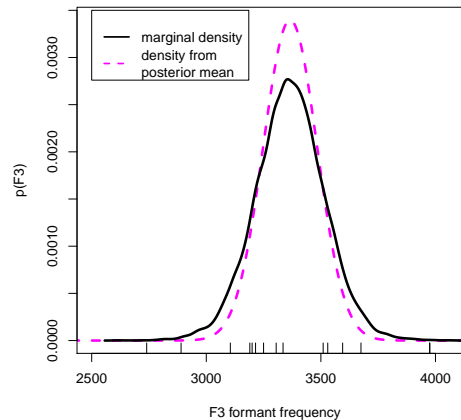
```
> update(m, 1000)
> res <- coda.samples(m, c("mu", "sigma", "predictions"), n.iter = 20000,
+   thin = 1)
```

and extract the relevant statistics and plot the outcome as follows:

```
> mu.mean <- mean(res[[1]][, 1])
> sigma.mean <- mean(res[[1]][, 12])
> from <- 1800
> to <- 4800
> x <- seq(from, to, by = 1)
> plot(x, dnorm(x, mu.mean, sigma.mean), col = "magenta", lwd = 3,
+   lty = 2, type = "l", xlim = c(from, to), xlab = "F3 formant frequency",
+   ylab = "p(F3)")
```



(a) Estimated density



(b) Density estimate over mean of ten new observations

Figure 4.12: Bayesian inference for normal distribution

```
> lines(density(res[[1]][, 2], from = from, to = to), lwd = 3)
> rug(response)
> legend(from, 0.0011, c("marginal density", "density from\nposterior mean"),
+       lty = c(1, 2), lwd = 2, col = c("black", "magenta"))
> from <- 2500
> to <- 4100
> plot(x, dnorm(x, mu.mean, sigma.mean/sqrt(M)), type = "l", lty = 2,
+      col = "magenta", lwd = 3, xlim = c(from, to), xlab = "F3 formant frequency",
+      ylab = "p(F3)")
> lines(density(apply(res[[1]][, 2:11], 1, mean, from = from, to = to)),
+      lwd = 3)
> rug(response)
> legend(from, 0.0035, c("marginal density", "density from\nposterior mean"),
+       lty = c(1, 2), lwd = 2, col = c("black", "magenta"))
```

The resulting density estimate for a single future observation is shown in Figure 4.12a. This is almost the same as the result obtained from using the posterior mean. However, the density estimate for the mean obtained in ten future observations, shown in Figure 4.12b, is rather different: once again it has greater dispersion than the estimate obtained using the posterior mean.<sup>13</sup>

The ability to specify model structures like this, drawing from a variety of distributions, and to compute approximate posterior densities with general-purpose tools, gives tremen-

<sup>13</sup>The density on the mean of ten future observations under the posterior mean  $\mu$  and  $\sigma^2$  is given by expressing the mean as a linear combination of ten independent identically distributed normal random variables (Section 3.3).

dous modeling flexibility. The only real limits are conceptual—coming up with probabilistic models that are appropriate for a given type of data—and computational—time and memory.

## 4.6 Further reading

Gelman et al. (2004) is probably the best reference for practical details and advice in Bayesian parameter estimation and prediction.

## 4.7 Exercises

### Exercise 4.1

Confirm using simulations that the variance of relative-frequency estimation of  $\pi$  for binomially distributed data really is  $\frac{\pi(1-\pi)}{n}$ : for all possible combinations of  $\pi \in \{0.1, 0.2, 0.5\}$ ,  $n \in \{10, 100, 1000\}$ , randomly generate 1000 datasets and estimate  $\hat{\pi}$  using relative frequency estimation. Plot the observed variance against the variance predicted in Equation 4.1

### Exercise 4.2: Maximum-likelihood estimation for the geometric distribution

You encountered the geometric distribution in Chapter 3, which models the generation of sequence lengths as the repeated flipping of a weighted coin until a single success is achieved. Its lone parameter is the success parameter  $\pi$ . Suppose that you have a set of observed sequence lengths  $\mathbf{y} = y_1, \dots, y_n$ . Since a sequence of length  $k$  corresponds to  $k - 1$  “failures” and one “success”, the total number of “failures” in  $\mathbf{y}$  is  $\sum_i (y_i - 1)$  and the total number of “successes” is  $n$ .

1. From analogy to the binomial distribution, guess the maximum-likelihood estimate of  $\pi$ .
2. Is your guess of the maximum-likelihood estimate biased? You’re welcome to answer this question either through mathematical analysis or through computational simulation (i.e. choose a value of  $\pi$ , repeatedly generate sets of geometrically-distributed sequences using your choice of  $\pi$ , and quantify the discrepancy between the average estimate  $\hat{\pi}$  and the true value).
3. Use your estimator to find best-fit distributions for token-frequency and type-frequency distributions of word length in syllables as found in the file `brown-counts-lengths-nsyll` (parsed Brown corpus; see Exercise 3.6).

### Exercise 4.3

We covered Bayesian parameter estimation for the binomial distribution where the prior distribution on the binomial success parameter  $\pi$  was of the form

$$P(\pi) \propto \pi^a(1 - \pi)^b$$

Plot the shape of this prior for a variety of choices of  $a$  and  $b$ . What determines the mode of the distribution (i.e., the value of  $\pi$  where the curve's maximum lies) and its degree of peakedness? What do  $a$  and  $b$  together represent?

#### **Exercise 4.4: Binomial and beta-binomial predictive distributions**

Three native English speakers start studying a new language together. This language has flexible word order, so that sometimes the subject of the sentence can precede the verb (SV), and sometimes it can follow the verb (VS). Of the first three utterances of the new language they are taught, one is VS and the other two are SV.

Speaker A abandons her English-language preconceptions and uses the method of maximum likelihood to estimate the probability that an utterance will be SV. Speakers B and C carry over some preconceptions from English; they draw inferences regarding the SV/VS word order frequency in the language according to a beta-distributed prior, with  $\alpha_1 = 8$  and  $\alpha_2 = 1$  (here, SV word order counts as a “success”). Speaker B uses maximum a-posterior (MAP) probability to estimate the probability that an utterance will be SV. Speaker C is fully Bayesian and retains a full posterior distribution on the probability that an utterance will be SV.

It turns out that the first three utterances of the new language were uncharacteristic; of the next twenty-four utterances our speakers hear, sixteen of them are VS. Which of our three speakers was best prepared for this eventuality, as judged by the predictive distribution placed by the speaker on the word order outcomes of these twenty-four utterances? Which of our speakers was worst prepared? Why?

#### **Exercise 4.5: What level of autocorrelation is acceptable in a Markov chain?**

How do you know when a given level of autocorrelation in a thinned Markov chain is acceptably low? One way of thinking about this problem is to realize that a sequence of independent samples is generally going to have *some* non-zero autocorrelation, by pure chance. The longer such a sequence, however, the lower the autocorrelation is likely to be. (Why?) Simulate a number of such sequences of length  $N = 100$ , drawn from a uniform distribution, and compute the 97.5% quantile autocorrelation coefficient—that is, the value  $r$  such that 97.5% of the generated sequences have correlation coefficient smaller than this value. Now repeat this process for a number of different lengths  $N$ , and plot this threshold  $r$  as a function of  $N$ .

#### **Exercise 4.6: Autocorrelation of Markov-chain samples from BUGS.**

Explore the autocorrelation of the samples obtained in the two models of Section 4.5, varying how densely you subsample the Markov chain by varying the thinning interval (specified by the `thin` argument of `coda.samples()`). Plot the average (over 20 runs) autocorrelation on each model parameter as a function of the thinning interval. For each model, how sparsely do you need to subsample the chain in order to effectively eliminate the autocorrelation? **Hint:** in R, you can compute the autocorrelation of a vector  $\mathbf{x}$  with:

```
> cor(x[-1], x[-length(x)])
```



# Chapter 5

## Confidence Intervals and Hypothesis Testing

Although Chapter 4 introduced the theoretical framework for estimating the parameters of a model, it was very much situated in the context of prediction: the focus of statistical inference is on inferring the kinds of additional data that are likely to be generated by a model, on the basis of an existing set of observations. In much of scientific inquiry, however, we wish to use data to make inferences about models themselves: what plausible range can be inferred for a parameter or set of parameters within a model, or which of multiple models a given set of data most strongly supports. These are the problems of CONFIDENCE INTERVALS and HYPOTHESIS TESTING respectively. This chapter covers the fundamentals of Bayesian and frequentist approaches to these problems.

### 5.1 Bayesian confidence intervals

Recall from Section 4.4 that Bayesian parameter estimation simply involves placing a posterior probability distribution over the parameters  $\theta$  of a model, on the basis of Bayes rule:

$$P(\theta|\mathbf{y}) = \frac{P(\mathbf{y}|\theta)P(\theta)}{P(\mathbf{y})} \quad (5.1)$$

In Bayesian inference, a CONFIDENCE INTERVAL over a single model parameter  $\phi$  is simply a contiguous interval  $[\phi_1, \phi_2]$  that contains a specified proportion of the posterior probability mass over  $\phi$ . The proportion of probability mass contained in the confidence interval can be chosen depending on whether one wants a narrower or wider interval. The tightness of the interval (in frequentist as well as Bayesian statistics) is denoted by a value  $\alpha$  that expresses the amount of probability mass *excluded* from the interval—so that  $(1 - \alpha)\%$  of the probability mass is within the interval. The interpretation of a  $(1 - \alpha)\%$  confidence interval  $[\phi_1, \phi_2]$  is that **the probability that the model parameter  $\phi$  resides in  $[\phi_1, \phi_2]$  is  $(1 - \alpha)$ .**

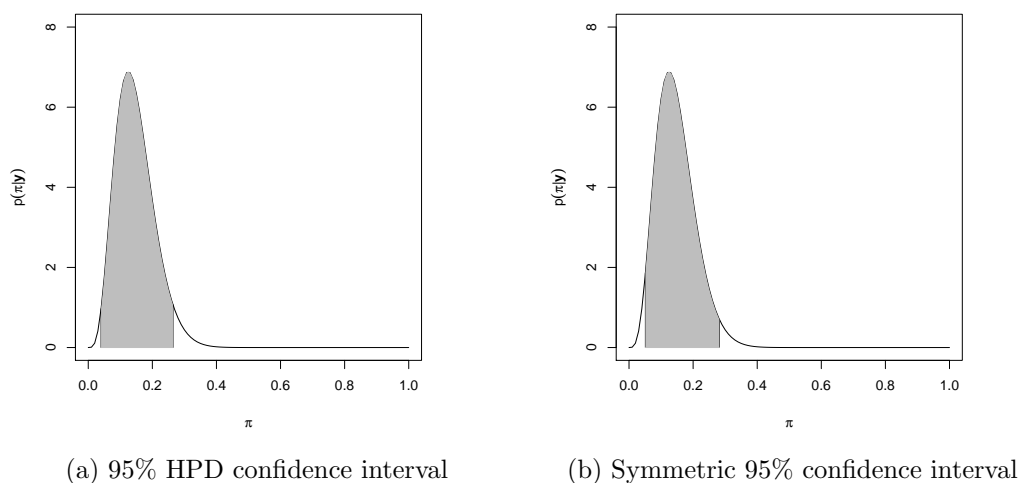


Figure 5.1: HPD and symmetric Bayesian confidence intervals for a posterior distributed as  $Beta(5, 29)$

Of course, there is always more than one way of choosing the bounds of the interval  $[\phi_1, \phi_2]$  to enclose  $(1 - \alpha)\%$  of the posterior mass. There are two main conventions for determining how to choose interval boundaries:

- Choose the *shortest* possible interval enclosing  $(1 - \alpha)\%$  of the posterior mass. This is called a HIGHEST POSTERIOR DENSITY (HPD) confidence interval.
- Choose interval boundaries such that an *equal amount of probability mass* is contained on either side of the interval. That is, choose  $[\phi_1, \phi_2]$  such that  $P(\phi < \phi_1 | \mathbf{y}) = P(\phi > \phi_2 | \mathbf{y}) = \frac{\alpha}{2}$ . This is called a SYMMETRIC confidence interval.

Let us return, for example, to our American English speaker of Chapter 4, assuming that she models speaker choice in passivization as a binomial random variable (with passive voice being “success”) with parameter  $\pi$  over which she has a Beta prior distribution with parameters  $(3, 24)$ , and observes five active and two passive clauses. The posterior over  $\pi$  has distribution  $Beta(5, 29)$ . Figure 5.1 shows HPD and symmetric 95% confidence intervals over  $\pi$ , shaded in gray, for this posterior distribution. The posterior is quite asymmetric, and for the HPD interval there is more probability mass to the right of the interval than there is to the left. The intervals themselves are, of course, qualitatively quite similar.

## 5.2 Bayesian hypothesis testing

In all types of statistics, hypothesis testing involves entertaining multiple candidate generative models of how observed data has been generated. The hypothesis test involves an assessment of which model is most strongly warranted by the data. Bayesian hypothesis

testing in particular works just like any other type of Bayesian inference. Suppose that we have a collection of hypotheses  $H_1, \dots, H_n$ . Informally, a hypothesis can range over diverse ideas such as “this coin is fair”, “the animacy of the agent of a clause affects the tendency of speakers to use the passive voice”, “females have higher average F1 vowel formants than males regardless of the specific vowel”, or “a word’s frequency has no effect on naming latency”. Formally, each hypothesis should specify a model that determines a probability distribution over possible observations  $\mathbf{y}$ . Furthermore, we need a prior probability over the collection of hypotheses,  $P(H_i)$ . Once we have observed some data  $\mathbf{y}$ , we use Bayes’ rule (Section 2.4.1) to calculate the posterior probability distribution over hypotheses:

$$P(H_i|\mathbf{y}) = \frac{P(\mathbf{y}|H_i)P(H_i)}{P(\mathbf{y})} \quad (5.2)$$

where  $P(\mathbf{y})$  marginalizes (Section 3.2) over the hypotheses:

$$P(\mathbf{y}) = \sum_{j=1}^n P(\mathbf{y}|H_j)P(H_j) \quad (5.3)$$

As an example, let us return once more to the case of English binomials, such as *salt and pepper*. A number of constraints have been hypothesized to play a role in determining binomial ordering preferences; as an example, one hypothesized constraint is that ordered binomials of the form *A and B* should be disfavored when *B* has ultimate-syllable stress (\*BSTR; Bolinger, 1962; Müller, 1997). For example, *pepper and salt* violates this constraint against ultimate-syllable stress, but its alternate *salt and pepper* does not. We can construct a simple probabilistic model of the role of \*BSTR in binomial ordering preferences by assuming that every time an English binomial is produced that could potentially violate \*BSTR, the binomial is produced in the satisfying order *B and A* ordering with probability  $\pi$ , otherwise it is produced in the violating ordering *A and B*.<sup>1</sup> If we observe  $n$  such English binomials, then the distribution over the number of satisfactions of \*BSTR observed is (appropriately enough) the binomial distribution with parameters  $\pi$  and  $n$ .

Let us now entertain two hypotheses about the possible role of \*BSTR in determining binomial ordering preferences. In the first hypothesis,  $H_1$ , \*BSTR plays no role, hence orderings *A and B* and *B and A* are equally probable; we call this the “no-preference” hypothesis. Therefore in  $H_1$  the binomial parameter  $\pi$  is 0.5. In Bayesian inference, we need to assign probability distributions to choices for model parameters, so we state  $H_1$  as:

$$H_1 : P(\pi|H_1) = \begin{cases} 1 & \pi = 0.5 \\ 0 & \pi \neq 0.5 \end{cases}$$

The probability above is a PRIOR PROBABILITY on the binomial parameter  $\pi$ .

---

<sup>1</sup>For now we ignore the role of multiple overlapping constraints in jointly determining ordering preferences, as well as the fact that specific binomials may have idiosyncratic ordering preferences above and beyond their constituent constraints. The tools to deal with these factors are introduced in Chapters 6 and 8 respectively.

In our second hypothesis  $H_2$ , \*BSTR *does* affect binomial ordering preferences (the “preference” hypothesis). For this hypothesis we must place a non-trivial probability distribution on  $\pi$ . Keep in mind we have arbitrarily associated the “success” outcome with satisfaction of \*BSTR. Suppose that we consider only two possibilities in  $H_2$ : that the preference is either  $\frac{2}{3}$  for *A and B* or  $\frac{2}{3}$  for outcome *B and A*, and let these two preferences be equally likely in  $H_2$ . This gives us:

$$H_2 : P(\pi|H_2) = \begin{cases} 0.5 & \pi = \frac{1}{3} \\ 0.5 & \pi = \frac{2}{3} \end{cases} \quad (5.4)$$

In order to complete the Bayesian inference of Equation (5.2), we need prior probabilities on the hypotheses themselves,  $P(H_1)$  and  $P(H_2)$ . If we had strong beliefs one way or another about the binomial’s ordering preference (e.g., from prior experience with other English binomials, or with experience with a semantically equivalent binomial in other languages), we might set one of these prior probabilities close to 1. For these purposes, we will use  $P(H_1) = P(H_2) = 0.5$ .

Now suppose we collect a dataset  $\mathbf{y}$  of six English binomials in which two orderings violate \*BSTR from a corpus:

| Binomial                       | Constraint status (S: *BSTR satisfied, V: *BSTR violated) |
|--------------------------------|---|
| <i>salt and pepper</i>         | S   |
| <i>complete and unabridged</i> | S   |
| <i>scabrous and unclean</i>    | V   |
| <i>grow and produce</i>        | S   |
| <i>ungallant and untrue</i>    | V   |
| <i>bold and entertaining</i>   | S   |

Do these data favor  $H_1$  or  $H_2$ ?

We answer this question by completing Equation (5.2). We have:

$$P(H_1) = 0.5$$

$$P(\mathbf{y}|H_1) = \binom{6}{4} \pi^4(1 - \pi)^2 = \binom{6}{4} \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^2 = 0.23$$

Now to complete the calculation of  $P(\mathbf{y})$  in Equation (5.3), we need  $P(\mathbf{y}|H_2)$ . To get this, we need to marginalize over the possible values of  $\pi$ , just as we are marginalizing over  $H$  to get the probability of the data. We have:

$$\begin{aligned}
P(\mathbf{y}|H_2) &= \sum_i P(\mathbf{y}|\pi_i) P(\pi_i|H_2) \\
&= P\left(\mathbf{y}|\pi = \frac{1}{3}\right) P\left(\pi = \frac{1}{3}|H_2\right) + P\left(\mathbf{y}|\pi = \frac{2}{3}\right) P\left(\pi = \frac{2}{3}|H_2\right) \\
&= \binom{6}{4} \left(\frac{1}{3}\right)^4 \left(\frac{2}{3}\right)^2 \times 0.5 + \binom{6}{4} \left(\frac{2}{3}\right)^4 \left(\frac{1}{3}\right)^2 \times 0.5 \\
&= 0.21
\end{aligned}$$

thus

$$P(\mathbf{y}) = \underbrace{0.23}_{P(\mathbf{y}|H_1)} \times \underbrace{0.5}_{P(H_1)} + \underbrace{0.21}_{P(\mathbf{y}|H_2)} \times \underbrace{0.5}_{P(H_2)} \quad (5.5)$$

$$= 0.22 \quad (5.6)$$

And we have

$$P(H_1|\mathbf{y}) = \frac{0.23 \times 0.5}{0.22} \quad (5.7)$$

$$= 0.53 \quad (5.8)$$

Note that even though the maximum-likelihood estimate of  $\hat{\pi}$  from the data we observed is exactly one of the two possible values of  $\pi$  under  $H_2$ , our data in fact support the “preference” hypothesis  $H_1$  – it went from prior probability  $P(H_1) = 0.5$  up to posterior probability  $P(H_1|\mathbf{y}) = 0.53$ . See also Exercise 5.3.

### 5.2.1 More complex hypotheses

We might also want to consider more complex hypotheses than  $H_2$  above as the “preference” hypothesis. For example, we might think all possible values of  $\pi$  in  $[0, 1]$  are equally probable *a priori*:

$$H_3 : P(\pi|H_3) = 1 \quad 0 \leq \pi \leq 1$$

(In Hypothesis 3, the probability distribution over  $\pi$  is continuous, not discrete, so  $H_3$  is still a proper probability distribution.) Let us discard  $H_2$  and now compare  $H_1$  against  $H_3$ .

Let us compare  $H_3$  against  $H_1$  for the same data. To do so, we need to calculate the likelihood  $P(\mathbf{y}|H_3)$ , and to do this, we need to marginalize over  $\pi$ . Since  $\pi$  can take on a continuous range of values under  $H_3$ , this marginalization takes the form of an integral:

$$P(\mathbf{y}|H_3) = \binom{6}{4} \int_0^1 \pi^4 (1 - \pi)^2 d\pi$$

We use the critical trick of recognizing this integral as a beta function (Section 4.4.2), which gives us:

$$= \binom{6}{4} B(5, 3) = 0.14$$

If we plug this result back in, we find that

$$P(H_1|\mathbf{y}) = \frac{\overbrace{0.23}^{P(\mathbf{y}|H_1)} \times \overbrace{0.5}^{P(H_1)}}{\underbrace{0.23}_{P(\mathbf{y}|H_1)} \times \underbrace{0.5}_{P(H_1)} + \underbrace{0.14}_{P(\mathbf{y}|H_3)} \times \underbrace{0.5}_{P(H_3)}} = 0.62$$

So  $H_3$  fares even worse than  $H_2$  against the no-preference hypothesis  $H_1$ . Correspondingly, we would find that  $H_2$  is favored over  $H_3$ .

## 5.2.2 Bayes factor

Sometimes we do not have strong feelings about the prior probabilities  $P(H_i)$ . Nevertheless, we can quantify how much evidence a given dataset provides for one hypothesis over another. We can express the relative preference between  $H$  and  $H'$  in the face of data  $\mathbf{y}$  in terms of the PRIOR ODDS of  $H$  versus  $H'$  combined with the LIKELIHOOD RATIO between the two hypotheses. This combination gives us the POSTERIOR ODDS:

$$\frac{\overbrace{P(H|\mathbf{y})}^{\text{Posterior odds}}}{\overbrace{P(H'|\mathbf{y})}^{\text{Posterior odds}}} = \frac{\overbrace{P(\mathbf{y}|H)}^{\text{Likelihood ratio}}}{\overbrace{P(\mathbf{y}|H')}^{\text{Likelihood ratio}}} \frac{\overbrace{P(H)}^{\text{Prior odds}}}{\overbrace{P(H')}^{\text{Prior odds}}}$$

The contribution of the data  $\mathbf{y}$  to the posterior odds is simply the likelihood ratio:

$$\frac{P(\mathbf{y}|H)}{P(\mathbf{y}|H')} \tag{5.9}$$

which is also called the BAYES FACTOR between  $H$  and  $H'$ . A Bayes factor above 1 indicates support for  $H$  over  $H'$ ; a Bayes factor below 1 indicates support for  $H'$  over  $H$ . For example, the Bayes factors for  $H_1$  versus  $H_2$  and  $H_1$  versus  $H_3$  in the preceding examples

$$\frac{P(\mathbf{y}|H_1)}{P(\mathbf{y}|H_2)} = \frac{0.23}{0.21} = 1.14 \qquad \frac{P(\mathbf{y}|H_1)}{P(\mathbf{y}|H_3)} = \frac{0.23}{0.14} = 1.64$$

indicating weak support for  $H_1$  in both cases.

### 5.2.3 Example: Learning contextual contingencies in sequences

One of the key tasks of a language learner is to determine which cues to attend to in learning distributional facts of the language in their environment (Saffran et al., 1996a; Aslin et al., 1998; Swingley, 2005; Goldwater et al., 2007). In many cases, this problem of cue relevance can be framed in terms of hypothesis testing or model selection.

As a simplified example, consider a length-21 sequence of syllables:

da ta da ta ta da da da da da ta ta ta da ta ta ta da da da da

Let us entertain two hypotheses. The first hypothesis  $H_1$ , is that the probability of an **da** is independent of the context. The second hypothesis,  $H_2$ , is that the probability of an **da** is dependent on the preceding token. The learner’s problem is to choose between these hypotheses—that is, to decide whether immediately preceding context is relevant in estimating the probability distribution over what the next phoneme will be. How should the above data influence the learner’s choice? (Before proceeding, you might want to take a moment to examine the sequence carefully and answer this question on the basis of your own intuition.)

We can make these hypotheses precise in terms of the parameters that each entails.  $H_1$  involves only one binomial parameter  $P(\text{da})$ , which we will denote as  $\pi$ .  $H_2$  involves three binomial parameters:

1.  $P(\text{da}|\emptyset)$  (the probability that the sequence will start with **da**), which we will denote as  $\pi_\emptyset$ ;
2.  $P(\text{da}|\text{da})$  (the probability that an **da** will appear after an **da**), which we will denote as  $\pi_{\text{da}}$ ;
3.  $P(\text{da}|\text{ta})$  (the probability that an **da** will appear after an **ta**), which we will denote as  $\pi_{\text{ta}}$ .

(For expository purposes we will assume that the probability distribution over the number of syllables in the utterance is the same under both  $H_1$  and  $H_2$  and hence plays no role in the Bayes factor.) Let us assume that  $H_1$  and  $H_2$  are equally likely; we will be concerned with the Bayes factor between the two hypotheses. We will put a uniform prior distribution on all model parameters—recall that this can be expressed as a beta density with parameters  $\alpha_1 = \alpha_2 = 1$  (Section 4.4.2).

There are 21 observations, 12 of which are **da** and 9 of which are **ta**. The likelihood of  $H_1$  is therefore simply

$$\int_0^1 \pi^{12}(1 - \pi)^9 d\pi = B(13, 10) \\ = 1.55 \times 10^{-7}$$

once again recognizing the integral as a beta function (see Section 4.4.2).

To calculate the likelihood of  $H_2$  it helps to lay out the 21 events as a table of conditioning contexts and outcomes:

|             | Outcome |    |
|-------------|---------|----|
| Context     | da      | ta |
| $\emptyset$ | 1       | 0  |
| da          | 7       | 4  |
| da          | 4       | 5  |

The likelihood of  $H_2$  is therefore

$$\int_0^1 \pi_\emptyset^1 d\pi_\emptyset \int_0^1 \pi_{\text{da}}^7 (1 - \pi_{\text{da}})^4 d\pi_{\text{da}} \int_0^1 \pi_{\text{ta}}^4 (1 - \pi_{\text{ta}})^5 d\pi_{\text{ta}} = B(2, 1)B(8, 5)B(5, 6) \\ = 1 \times 10^{-7}$$

This dataset provides some support for the simpler hypothesis of statistical independence—the Bayes factor is 1.55 in favor of  $H_1$ .

### 5.3 Frequentist confidence intervals

We now move on to frequentist confidence intervals and hypothesis testing, which have been developed from a different philosophical standpoint. To a frequentist, it does not make sense to say that “the true parameter  $\theta$  lies between these points  $x$  and  $y$  with probability  $p^*$ .” The parameter  $\theta$  is a real property of the population from which the sample was obtained and is either in between  $x$  and  $y$ , or it is not. Remember, to a frequentist, the notion of probability as reasonable belief is not admitted! Under this perspective, the Bayesian definition of a confidence interval—while intuitively appealing to many—is incoherent.

Instead, the frequentist uses more indirect means of quantifying their certainty about the estimate of  $\theta$ . The issue is phrased thus: imagine that I were to repeat the same experiment—drawing a sample from my population—many times, and each time I repeated the experiment I constructed an interval  $I$  on the basis of my sample according to a fixed procedure **Proc**. Suppose it were the case that  $p$  percent of the intervals  $I$  thus constructed actually contained  $\theta$ . Then for any given sample  $S$ , the interval  $I$  constructed by **Proc** is a  $(1 - p)\%$  confidence interval for  $\theta$ .

If you think that this seems like convoluted logic, well, you are not alone. **Frequentist confidence intervals are one of the most widely misunderstood constructs in statistics.** The Bayesian view is more intuitive to most people. Under some circumstances, there is a happy coincidence where Bayesian and frequentist confidence intervals look the same and you are free to misinterpret the latter as the former. In general, however, they do *not* necessarily look the same, and you need to be careful to interpret each correctly.



Here’s an example, where we will explain the STANDARD ERROR OF THE MEAN. Suppose that we obtain a sample of  $n$  observations from a normal distribution  $N(\mu, \sigma^2)$ . It turns out that the following quantity follows the  $t_{n-1}$  distribution (Section B.5):

$$\frac{\hat{\mu} - \mu}{\sqrt{S^2/n}} \sim t_{n-1} \quad (5.10)$$

where

$$\hat{\mu} = \frac{1}{n} \sum_i X_i \quad \text{[maximum-likelihood estimate of the mean]}$$

$$S^2 = \frac{1}{n-1} \sum_i (X_i - \hat{\mu})^2 \quad \text{[unbiased estimate of } \sigma^2 \text{; Section 4.3.3]}$$

Let us denote the quantile function for the  $t_{n-1}$  distribution as  $Q_{t_{n-1}}$ . We want to choose a symmetric interval  $[-a, a]$  containing  $(1 - \alpha)$  of the probability mass of  $t_{n-1}$ . Since the  $t$  distribution is symmetric around 0, if we set  $a = \sqrt{S^2/n} Q_{t_{n-1}}(1 - \alpha/2)$ , we will have

$$P(\hat{\mu} - \mu < -a) = \frac{\alpha}{2} \quad (5.11)$$

$$P(\hat{\mu} - \mu > a) = \frac{\alpha}{2}$$

Figure 5.2 illustrates this for  $\alpha = 0.05$  (a 95% confidence interval). Most of the time, the “standardized” difference between  $\hat{\mu}$  and  $\mu$  is small and falls in the unshaded area. But 5% of the time, this standardized difference will fall in the shaded area—that is, the confidence interval won’t contain  $\mu$ .

Note that the quantity  $S/\sqrt{n}$  is called the the STANDARD ERROR OF THE MEAN or simply the STANDARD ERROR. Note that this is different from the standard deviation of the sample, but related! (How?) When the number of observations  $n$  is large, the  $t$  distribution looks approximately normal, and as a rule of thumb, the symmetric 95% tail region of the normal distribution is about 2 standard errors away from the mean.

Another example: let’s look at the data from a classic study of the English vowel space (Peterson and Barney, 1952). The distribution of the F1 formant for the vowel  $\varepsilon$  is roughly normal (see Figure 5.3). The 95% confidence interval can be calculated by looking at the quantity  $S/\sqrt{n} Q_{t_{151}}(0.975) = 15.6$ . This is half the length of the confidence interval; the confidence interval should be centered around the sample mean  $\hat{\mu} = 590.7$ . Therefore our 95% confidence interval for the mean F1 is  $[575.1, 606.3]$ .

## 5.4 Frequentist hypothesis testing

In most of science, including areas such as psycholinguistics and phonetics, statistical inference is most often seen in the form of hypothesis testing within the NEYMAN-PEARSON

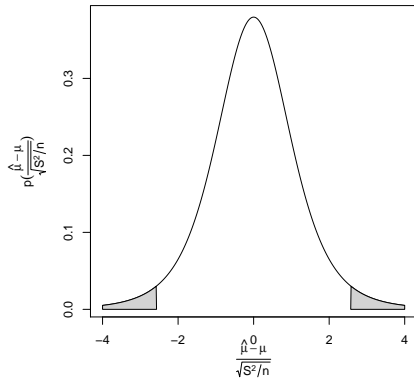


Figure 5.2: Visualizing confidence intervals with the  $t$  distribution

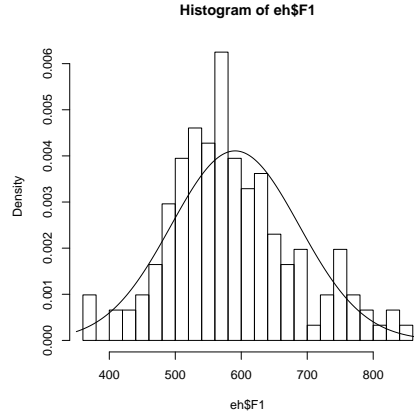


Figure 5.3: Distribution of F1 formant for  $[\epsilon]$

PARADIGM. This paradigm involves formulating two hypotheses, the NULL HYPOTHESIS  $H_0$  and a more general ALTERNATIVE HYPOTHESIS  $H_A$  (sometimes denoted  $H_1$ ). We then design a *decision procedure* which involves collecting some data  $\mathbf{y}$  and computing a statistic  $T(\mathbf{y})$ , or just  $T$  for short. Before collecting the data  $\mathbf{y}$ ,  $T(\mathbf{y})$  is a random variable, though we do not know its distribution because we do not know whether  $H_0$  is true. At this point we divide the range of possible values of  $T$  into an ACCEPTANCE REGION and a REJECTION REGION. Once we collect the data, we accept the null hypothesis  $H_0$  if  $T$  falls into the acceptance region, and reject  $H_0$  if  $T$  falls into the rejection region.

Now,  $T$  is a random variable that will have one distribution under  $H_0$ , and another distribution under  $H_A$ . Let us denote the probability mass in the rejection region under  $H_0$  as  $\alpha$ , and the mass in the same region under  $H_A$  as  $1 - \beta$ . There are four logically possible combinations of the truth value of  $H_0$  and our decision once we have collected  $\mathbf{y}$ :

|                 |       |  |                                       |
|-----------------|-------|--|---------------------------------------|
|                 |       | Our decision                           |                                       |
|                 |       | Accept $H_0$                           | Reject $H_0$                          |
| (1) $H_0$ is... | True  | Correct decision (prob. $1 - \alpha$ ) | Type I error (prob. $\alpha$ )        |
|                 | False | Type II error (prob. $\beta$ )         | Correct decision (prob. $1 - \beta$ ) |

The probabilities in each row of I sum to 1, since they represent the conditional probability of our decision given the truth/falsity of  $H_0$ .

As you can see in I, there are two sets of circumstances under which we have done the right thing:

1. The null hypothesis is true, and we accept it (probability  $1 - \alpha$ ).
2. The null hypothesis is false, and we reject it (probability  $1 - \beta$ ).

This leaves us with two sets of circumstances under which we have made an error:

1. The null hypothesis is true, but we reject it (probability  $\alpha$ ). This by convention is called a **TYPE I ERROR**.
2. The null hypothesis is false, but we accept it (probability  $\beta$ ). This by convention is called a **TYPE II ERROR**.

For example, suppose that a psycholinguist uses a simple visual world paradigm to examine the time course of word recognition. She presents to participants a display on which a desk is depicted on the left, and a duck is depicted on the right. Participants start with their gaze on the center of the screen, and their eye movements are recorded as they hear the word “duck”. The question at issue is whether participants’ eye gaze fall reliably more often on the duck than on the desk in the window 200 – 250 milliseconds after the onset of “duck”, and the researcher devises a simple rule of thumb that if there are more than twice as many fixations on the duck than on the chair within this window, the null hypothesis will be rejected. Her experimental results involve 21% fixations on the duck and 9% fixations on the chair, so she rejects the null hypothesis. However, she later finds out that her computer was miscalibrated by 300 milliseconds and the participants had not even heard the onset of the word by the end of the relevant window. The researcher had committed a Type I error. (In this type of scenario, a Type I error is often called a **FALSE POSITIVE**, and a Type II error is often called a **FALSE NEGATIVE**.)

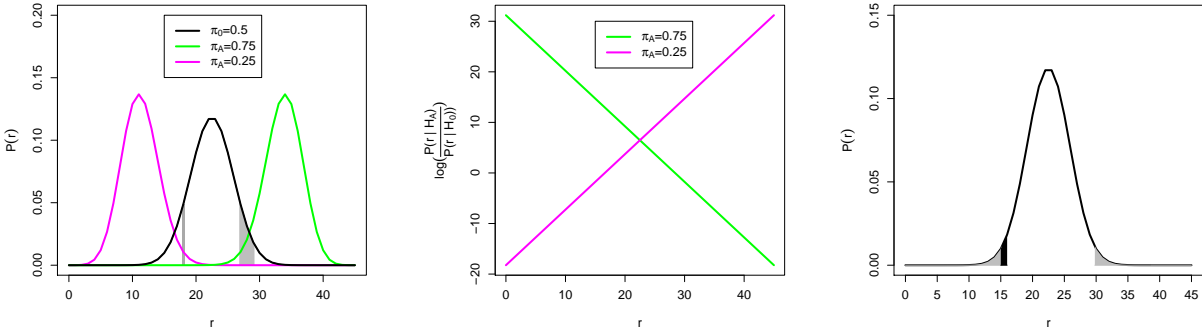
The probability  $\alpha$  of Type I error is referred to as the **SIGNIFICANCE LEVEL** of the hypothesis test. In the Neyman-Pearson paradigm,  $T$  is always chosen such that its (asymptotic) distribution can be computed. The probability  $1 - \beta$  of *not* committing Type II error is called the **POWER** of the hypothesis test. There is always a trade-off between significance level and power, but the goal is to use decision procedures that have the highest possible power for a given significance level. To calculate  $\beta$  and thus the power, however, we need to know the true model, so determining the optimality of a decision procedure with respect to power can be tricky.

Now we’ll move on to a concrete example of hypothesis testing in which we deploy some probability theory.

### 5.4.1 Hypothesis testing: binomial ordering preference

One of the simplest cases of hypothesis testing—and one that is often useful in the study of language—is the **BINOMIAL TEST**, which we illustrate here.

You decide to investigate the role of ultimate-syllable stress avoidance in English binomial ordering preferences by collecting from the British National Corpus 45 tokens of binomials in which \*BSTR could be violated. As the test statistic  $T$  you simply choose the number of successes  $r$  in these 45 tokens. Therefore the distribution of  $T$  under the no-preference null hypothesis  $H_0 : \pi = 0.5$  is simply the distribution on the number of successes  $r$  for a binomial distribution with parameters  $n = 45, \pi = 0.5$ . The most general natural alternative hypothesis  $H_A$  of “preference” would be that the binomial has some arbitrary preference



(a) Probability densities over  $r$  for  $H_0$  and two instantiations of  $H_A$  (b) The log-ratio of probability densities under  $H_0$  and two instantiations of  $H_A$  (c) Standard (maximum-power) two-tailed and one-tailed rejection regions

Figure 5.4: The trade-off between significance level and power

$$H_A : 0 \leq \pi \leq 1 \tag{5.12}$$

Unlike the case with Bayesian hypothesis testing, we do not put a probability distribution on  $\pi$  under  $H_A$ . We complete our decision procedure by partitioning the possible values of  $T$  into acceptance and rejection regions. To achieve a significance level  $\alpha$  we must choose a partitioning such that the rejection region contains probability mass of no more than  $\alpha$  under the null hypothesis. There are many such partitions that achieve this. For example, the probability of achieving 18 successes in 45 trials is just under 5%; so is the probability of achieving at least 27 successes but not more than 29 successes. The black line in Figure 5.4a shows the probability density function for  $H_0$ , and each of the gray areas corresponds to one of these rejection regions.

However, the principle of maximizing statistical power helps us out here. Recall that when  $H_A$  is true, the power of the hypothesis test,  $1 - \beta$ , is the probability that  $T(\mathbf{y})$  will fall in our rejection region. The significance level  $\alpha$  that we want to achieve, however, constrains how large our rejection region can be. To maximize the power, it therefore makes sense to choose as the rejection region that part of the range of  $T$  assigned lowest probability by  $H_0$  and highest probability by  $H_A$ . Let us denote the probability mass functions for  $T$  under  $H_0$  and  $H_A$  as  $P_0(T)$  and  $P_A(T)$  respectively. Figure 5.4b illustrates the tradeoff by plotting the log-ratio  $\log \frac{P_A(T)}{P_0(T)}$  under two example instantiations of  $H_A$ :  $\pi_A = 0.25$  and  $\pi_A = 0.75$ . The larger this ratio for a given possible outcome  $t$  of  $T$ , the more power is obtained by inclusion of  $t$  in the rejection region. When  $\pi_A < 0.5$ , the most power is obtained by filling the rejection region with the largest possible values of  $T$ . Likewise, when  $\pi_A > 0.5$ , the most power is obtained by filling the rejection region with the smallest possible values of  $T$ .<sup>2</sup> Since our  $H_A$  entertains all possible values of  $\pi$ , we obtain maximal power by splitting our

<sup>2</sup>Although it is beyond the scope of this text to demonstrate it, this principle of maximizing statistical

rejection region into two symmetric halves, one on the left periphery and the other on the right periphery. In Figure 5.4c, the gray shaded area represents the largest such split region that contains less than 5% of the probability mass under  $H_0$  (actually  $\alpha = 0.0357$ ). If our 45 tokens do not result in at least 16 and at most 29 successes, we will reject  $H_0$  in favor of  $H_A$  under this decision rule. This type of rule is called a TWO-TAILED TEST because the rejection region is split equally in the two tails of the distribution of  $T$  under  $H_0$ .

Another common type of alternative hypothesis would be that is a tendency to *satisfy* \*BSTR. This alternative hypothesis would naturally be formulated as  $H'_A : 0.5 < \pi \leq 1$ . This case corresponds to the green line in Figure 5.4b; in this case we get the most power by putting our rejection region entirely on the left. The largest possible such rejection region for our example consists of the lefthand gray region plus the black region in Figure 5.4c ( $\alpha = 0.0362$ ). This is called a ONE-TAILED TEST. The common principle which derived the form of the one-tailed and two-tailed tests alike is the idea that *one should choose the rejection region that maximizes the power of the hypothesis test if  $H_A$  is true*.

Finally, a common approach in science is not simply to choose a significance level  $\alpha$  in advance and then report whether  $H_0$  is accepted or rejected, but rather to report the lowest value of  $\alpha$  for which  $H_0$  would be rejected. This is what is known as the *p*-VALUE of the hypothesis test. For example, if our 45 tokens resulted in 14 successes and we conducted a two-tailed test, we would compute twice the cumulative distribution function of Binom(45,0.5) at 14, which would give us an outcome of  $p = 0.016$ .

## 5.4.2 Quantifying strength of association in categorical variables

There are many situations in quantitative linguistic analysis where you will be interested in the possibility of association between two categorical variables. In this case, you will often want to represent your data as a contingency table. A  $2 \times 2$  contingency table has the following form:

$$\begin{array}{cc}
 & Y \\
 & \begin{array}{cc} y_1 & y_2 \end{array} \\
 X & \begin{array}{cc} x_1 & \left[ \begin{array}{cc} n_{11} & n_{12} \end{array} \right] n_{1*} \\
 & \begin{array}{cc} x_2 & \left[ \begin{array}{cc} n_{21} & n_{22} \end{array} \right] n_{2*} \\
 & \left[ \begin{array}{cc} n_{*1} & n_{*2} \end{array} \right] n_{**} \end{array} \end{array} \quad (5.13)
 \end{array}$$

where the  $n_{i*}$  are the marginal totals for different values of  $x_i$  across values of  $Y$ , the  $n_{*j}$  are the marginal totals for different values of  $y_j$  across values of  $X$ , and  $n_{**}$  is the grand total number of observations.

We'll illustrate the use of contingency tables with an example of quantitative syntax: the study of coordination. In traditional generative grammar, rules licensing coordination had the general form

---

power leads to the same rule for constructing a rejection region regardless of the precise values entertained for  $\pi$  under  $H_A$ , so long as values both above and below 0.5 are entertained.

NP → NP Conj NP

or even

X → X Conj X

encoding the intuition that many things could be coordinated with each other, but at some level every coordination should be a “combination of like categories”, a constraint referred to as CONJOIN LIKES (Chomsky, 1965). However, this approach turned out to be of limited success in a categorical context, as demonstrated by clear violations of like-category constraints such as II below (Sag et al., 1985; Peterson, 1986):

- (2) Pat is *a Republican* and *proud of it* (coordination of noun phrase with adjective phrase)

However, the preference for coordination to be between like categories is certainly strong as a *statistical tendency* (Levy, 2002). This in turn raises the question of whether the preference for coordinated constituents to be similar to one another extends to a level more fine grained than gross category structure (Levy, 2002; Dubey et al., 2008). Consider, for example, the following four coordinate noun phrases (NPs):

| Example |  | NP1   | NP2   |
|---------|--|-------|-------|
| 1.      | The girl and the boy                         | noPP  | noPP  |
| 2.      | [The girl from Quebec] and the boy           | hasPP | noPP  |
| 3.      | The girl and [the boy from Ottawa]           | noPP  | hasPP |
| 4.      | The girl from Quebec and the boy from Ottawa | hasPP | hasPP |

Versions 1 and 4 are *parallel* in the sense that both NP conjuncts have prepositional-phrase (PP) postmodifiers; versions 2 and 3 are non-parallel. If Conjoin Likes holds at the level of NP-internal PP postmodification as a violable preference, then we might expect coordinate NPs of types 1 and 4 to be more common than would “otherwise be expected”—a notion that can be made precise through the use of contingency tables.

For example, here are patterns of PP modifications in two-NP coordinations of this type from the parsed Brown and Switchboard corpora of English, expressed as 2 × 2 contingency tables:

|     |       | NP2   |      |             |     | NP2   |      |
|-----|-------|-------|------|-------------|-----|-------|------|
|     |       | hasPP | noPP |             |     | hasPP | noPP |
| (3) | Brown |       |      | Switchboard |     |       |      |
|     | NP1   | hasPP | noPP |             | NP1 | hasPP |      |
|     |       | 95    | 52   | 147         | 78  | 76    | 154  |
|     | noPP  | 174   | 946  | 1120        | 325 | 1230  | 1555 |
|     |       | 269   | 998  | 1267        | 403 | 1306  | 1709 |

From the table you can see that in both corpora, NP1 is more likely to have a PP postmodifier when NP2 has one, and NP2 is more likely to have a PP postmodifier when NP1 has one. But we would like to go beyond that and *quantify* the strength of the association between PP presence in NP1 on NP2. We would also like to *test for significance* of the association.

## Quantifying association: odds ratios

In Section 3.3 we already saw one method of quantifying the strength of association between two binary categorical variables: COVARIANCE or CORRELATION. Another popular way way of quantifying the predictive power of a binary variable  $X$  on another binary variable  $Y$  is with the ODDS RATIO. To introduce this concept, we first introduce the overall ODDS  $\omega^Y$  of  $y_1$  versus  $y_2$ :

mn

$$\omega^Y \stackrel{\text{def}}{=} \frac{n_{*1}}{n_{*2}} \quad (5.14)$$

Likewise, the odds  $\omega^X$  of  $x_1$  versus  $x_2$  are  $\frac{n_{1*}}{n_{2*}}$ . For example, in our Brown corpus examples we have  $\omega^Y = \frac{147}{1120} = 0.13$  and  $\omega^X = \frac{269}{998} = 0.27$ .

We further define the odds for  $Y$  if  $X = x_1$  as  $\omega_1^Y$  and so forth, giving us:

$$\omega_1^Y \stackrel{\text{def}}{=} \frac{n_{11}}{n_{12}} \quad \omega_2^Y \stackrel{\text{def}}{=} \frac{n_{21}}{n_{22}} \quad \omega_1^X \stackrel{\text{def}}{=} \frac{n_{11}}{n_{21}} \quad \omega_2^X \stackrel{\text{def}}{=} \frac{n_{12}}{n_{22}}$$

If the odds of  $Y$  for  $X = x_2$  are greater than the odds of  $Y$  for  $X = x_1$ , then the outcome of  $X = x_2$  **increases** the chances of  $Y = y_1$ . We can express the effect of the outcome of  $X$  on the odds of  $Y$  by the **odds ratio** (which turns out to be symmetric between  $X, Y$ ):

$$\mathcal{OR} \stackrel{\text{def}}{=} \frac{\omega_1^Y}{\omega_2^Y} = \frac{\omega_1^X}{\omega_2^X} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

An odds ratio  $\mathcal{OR} = 1$  indicates no association between the variables. For the Brown and Switchboard parallelism examples:

$$\mathcal{OR}_{\text{Brown}} = \frac{95 \times 946}{52 \times 174} = 9.93 \quad \mathcal{OR}_{\text{Swbd}} = \frac{78 \times 1230}{325 \times 76} = 3.88$$

So the presence of PPs in left and right conjunct NPs seems more strongly interconnected for the Brown (written) corpus than for the Switchboard (spoken). Intuitively, this difference might be interpreted as parallelism of PP presence/absence in NPs being an aspect of stylization that is stronger in written than in spoken language.

### 5.4.3 Testing significance of association in categorical variables

In frequentist statistics there are several ways to test the significance of the association between variables in a two-way contingency table. Although you may not be used to thinking about these tests as the comparison of two hypotheses in form of statistical models, they are!

### Fisher's exact test

Fisher's exact test applies to  $2 \times 2$  contingency tables such as (5.13). It takes as  $H_0$  the model in which all *marginal totals* are fixed, but that the individual cell totals are not—alternatively stated, that the individual outcomes of  $X$  and  $Y$  are independent. **This means that under  $H_0$ , the true underlying odds ratio  $\mathcal{OR}$  is 1.**  $H_A$  is the model that the individual outcomes of  $X$  and  $Y$  are not independent. With Fisher's exact test, the test statistic  $T$  is the odds ratio  $\mathcal{OR}$ , which follows the HYPERGEOMETRIC DISTRIBUTION under the null hypothesis (Section B.3).

An advantage of this test is that it computes the *exact*  $p$ -value (that is, the smallest  $\alpha$  for which  $H_0$  would be rejected). Because of this, Fisher's exact test can be used even for very small datasets. In contrast, many of the tests we cover elsewhere in this book (including the chi-squared and likelihood-ratio tests later in this section) compute  $p$ -values that are only asymptotically correct, and are unreliable for small datasets. As an example, consider the small hypothetical parallelism dataset given in IV below:

|     |     |       |      |     |    |
|-----|-----|-------|------|-----|----|
|     |     | NP2   |      |     |    |
|     |     | hasPP | noPP |     |    |
| (4) | NP1 | hasPP | 3    | 14  | 26 |
|     |     | noPP  | 22   | 61  | 83 |
|     |     | 34    | 75   | 109 |    |

The odds ratio is 2.36, and Fisher's exact test gives a  $p$ -value of 0.07. If we were to see twice the data in the exact same proportions, the odds ratio would stay the same, but the significance of Fisher's exact test for non-independence would increase.

### Chi-squared test

This is probably the best-known contingency-table test. It is very general and can be applied to arbitrary  $N$ -cell tables, if you have a model with  $k$  parameters that predicts expected values  $E_{ij}$  for all cells. For the chi-squared test, the test statistic is Pearson's  $X^2$ :

$$X^2 = \sum_{ij} \frac{[n_{ij} - E_{ij}]^2}{E_{ij}} \tag{5.15}$$

In the chi-squared test,  $H_A$  is the model that each cell in the table has its own parameter  $p_i$  in one big multinomial distribution. When the expected counts in each cell are large enough (the generally agreed lower bound is  $\geq 5$ ), the  $X^2$  statistic is approximately distributed as a chi-squared ( $\chi^2$ ) random variable with  $N - k - 1$  degrees of freedom (Section B.4). The  $\chi^2$  distribution is asymmetric and the rejection region is always placed in the right tail of the distribution (see Section B.4), so we can calculate the  $p$ -value by subtracting from one the value of the cumulative distribution function for the observed  $X^2$  test statistic.

The most common way of using Pearson's chi-squared test is to test for the independence of two factors in a two-way contingency table. Take a  $k \times l$  two-way table of the form:



|       |          |          |          |          |          |
|-------|----------|----------|----------|----------|----------|
|       | $y_1$    | $y_2$    | $\cdots$ | $y_l$    |          |
| $x_1$ | $n_{11}$ | $n_{12}$ | $\cdots$ | $n_{1l}$ | $n_{1*}$ |
| $x_2$ | $n_{21}$ | $n_{22}$ | $\cdots$ | $n_{2l}$ | $n_{2*}$ |
|       | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| $x_l$ | $n_{l1}$ | $n_{l2}$ | $\cdots$ | $n_{ll}$ | $n_{l*}$ |
|       | $n_{*1}$ | $n_{*2}$ | $\cdots$ | $n_{*l}$ | $n$      |

Our null hypothesis is that the  $x_i$  and  $y_i$  are independently distributed from one another. By the definition of probabilistic independence, that means that  $H_0$  is:

$$P(x_i, y_j) = P(x_i)P(y_j)$$

In the chi-squared test we use the relative-frequency (and hence maximum-likelihood; Section 4.3.1) estimates of the marginal probability that an observation will fall in each row or column:  $\hat{P}(x_i) = \frac{n_{i*}}{n}$  and  $\hat{P}(y_j) = \frac{n_{*j}}{n}$ . This gives us the formula for the expected counts in Equation (5.15):

$$E_{ij} = nP(x_i)P(y_j)$$

**Example:** For the Brown corpus data in III, we have

$$P(x_1) = \frac{147}{1267} = 0.1160 \qquad P(y_1) = \frac{269}{1267} = 0.2123 \qquad (5.16)$$

$$P(x_2) = \frac{1120}{1267} = 0.8840 \qquad P(y_2) = \frac{998}{1267} = 0.7877 \qquad (5.17)$$

giving us

$$E_{11} = 31.2 \quad E_{12} = 115.8 \qquad E_{21} = 237.8 \quad E_{22} = 882.2 \qquad (5.18)$$

Comparing with III, we get

$$X^2 = \frac{(95 - 31.2)^2}{31.2} + \frac{(52 - 115.8)^2}{115.8} + \frac{(174 - 237.8)^2}{237.8} + \frac{(946 - 882.2)^2}{882.2} \qquad (5.19)$$

$$= 187.3445 \qquad (5.20)$$

We had 2 parameters in our model of independence, and there are 4 cells, so  $X^2$  is distributed as  $\chi_1^2$  (since  $4 - 2 - 1 = 1$ ). The cumulative distribution function of  $\chi_1^2$  at 187.3 is essentially 1, so the  $p$ -value is vanishingly small; by any standards, the null hypothesis can be confidently rejected.

|  |        |       |       |          |
|--|--------|-------|-------|----------|
| <b>Example with larger data table:</b> |        | NP PP | NP NP | NP other |
|  | gave   | 17    | 79    | 34       |
|  | paid   | 14    | 4     | 9        |
|  | passed | 4     | 1     | 16       |

It is worth emphasizing, however, that the chi-squared test is not reliable when expected counts in some cells are very small. For the low-count table in IV, for example, the chi-squared test yields a significance level of  $p = 0.038$ . Fisher's exact test is the gold standard here, revealing that the chi-squared test is too aggressive in this case.

#### 5.4.4 Likelihood ratio test

With this test, the statistic you calculate for your data  $\mathbf{y}$  is the LIKELIHOOD RATIO

$$\Lambda^* = \frac{\max \text{Lik}_{H_0}(\mathbf{y})}{\max \text{Lik}_{H_A}(\mathbf{y})} \quad (5.21)$$

—that is, the ratio of the data likelihood under the MLE in  $H_0$  to the data likelihood under the MLE in  $H_A$ . This requires that you explicitly formulate  $H_0$  and  $H_A$ , since you need to find the MLEs and the corresponding likelihoods. The quantity

$$G^2 \stackrel{\text{def}}{=} -2 \log \Lambda^* \quad (5.22)$$

is sometimes called the DEVIANCE, and it is approximately chi-squared distributed (Section B.4) with degrees of freedom equal to the difference in the number of free parameters in  $H_A$  and  $H_0$ . (This test is also unreliable when expected cell counts are low, as in  $< 5$ .)

The likelihood-ratio test gives similar results to the chi-squared for contingency tables, but is more flexible because it allows the comparison of arbitrary nested models. We will see the likelihood-ratio test used repeatedly in later chapters.

**Example:** For the Brown corpus data above, let  $H_0$  be the model of independence between NP1 and NP2 with respective success parameters  $\pi_1$  and  $\pi_2$ , and  $H_A$  be the model of full non-independence, in which each complete outcome  $\langle x_i, y_j \rangle$  can have its own probability  $\pi_{ij}$  (this is sometimes called the SATURATED MODEL). We use maximum likelihood to fit each model, giving us for  $H_0$ :

$$\pi_1 = 0.116 \qquad \qquad \qquad \pi_2 = 0.212$$

and for  $H_A$ :

$$\pi_{11} = 0.075 \quad \pi_{12} = 0.041 \qquad \qquad \pi_{21} = 0.137 \quad \pi_{22} = 0.747$$

We calculate  $G^2$  as follows:

$$\begin{aligned} -2 \log \Lambda^* &= -2 \log \frac{(\pi_1 \pi_2)^{95} (\pi_1 (1 - \pi_2))^{52} ((1 - \pi_1) \pi_2)^{174} ((1 - \pi_1)(1 - \pi_2))^{946}}{\pi_{11}^{95} \pi_{12}^{52} \pi_{21}^{174} \pi_{22}^{946}} \\ &= -2 [95 \log(\pi_1 \pi_2) + 52 \log(\pi_1 (1 - \pi_2)) + 174 \log((1 - \pi_1) \pi_2) + 946 \log((1 - \pi_1)(1 - \pi_2))] \\ &\quad - 95 \pi_{11} - 52 \pi_{12} - 174 \pi_{21} - 946 \pi_{22}] \\ &= 151.6 \end{aligned}$$

$H_0$  has two free parameters, and  $H_A$  has three free parameters, so  $G^2$  should be approximately distributed as  $\chi_1^2$ . Once again, the cumulative distribution function of  $\chi_1^2$  at 151.6 is essentially 1, so the  $p$ -value of our hypothesis test is vanishingly small.

## 5.5 Exercises

### Exercise 5.1

Would the Bayesian hypothesis-testing results of Section 5.2 be changed at all if we did not consider the data as summarized by the number of successes and failures, and instead used the likelihood of the specific sequence HHTHTH instead? Why?

### Exercise 5.2

For Section 5.2, compute the posterior probabilities of  $H_1$ ,  $H_2$ , and  $H_3$  in a situation where all three hypotheses are entertained with prior probabilities  $P(H_1) = P(H_2) = P(H_3) = \frac{1}{3}$ .

### Exercise 5.3

Recompute the Bayesian hypothesis tests, computing both posterior probabilities and Bayes Factors, of Section 5.2 ( $H_1$  vs.  $H_2$  and  $H_1$  vs.  $H_3$ ) for the same data replicated twice – that is, the observations SSVSVSSSVSVS. Are the preferred hypotheses the same as for the original computations in Section 5.2? What about for the data replicated three times?

### Exercise 5.4

Use simulation to check that the theoretical confidence interval based on the  $t$  distribution for normally distributed data in Section 5.3 really works.

### Exercise 5.5

For a given choice of  $\alpha$ , is the procedure denoted in Equation (5.11) the only frequentist confidence interval that can be constructed for  $\mu$  for normally distributed data?

### Exercise 5.6

**Hypothesis testing: philosophy.** You surreptitiously observe an obsessed linguist compulsively searching a corpus for binomials of the form *pepper and salt* (P) or *salt and pepper* (S). He collects twenty examples, obtaining the sequence

SPSSSPSPSSSSPPSSSSSP

1. The linguist's research assistant tells you that the experiment was to obtain twenty examples and record the number of P's obtained. Can you reject the no-preference null hypothesis  $H_0$  at the  $\alpha = 0.05$  level?
2. The next day, the linguist tells you in class that she purposely misled the research assistant, and the actual experiment was to collect tokens from the corpus until six P examples were obtained and then stop. Does this new information affect the  $p$ -value with which you can reject the null hypothesis?

3. The linguist writes up her research results and sends them to a prestigious journal. The editor sends this article to two Bayesian reviewers. Both reviewers argue that this mode of hypothesis testing is ridiculous, and that a Bayesian hypothesis test should be made. Reviewer A suggests that the null hypothesis  $H_0$  of  $\pi = 0.5$  should be compared with the alternative hypothesis  $H_1$  of  $\pi = 0.25$ , and the two hypotheses should be given equal prior probability.
- (a) What is the posterior probability of  $H_0$  given the binomials data? Does the criteria by which the scientist decided how many binomials to collect affect the conclusions of a Bayesian hypothesis test? **Hint:** if  $\frac{P(H_0|\vec{x})}{P(H_1|\vec{x})} = a$ , then

$$P(H_0|\vec{x}) = \frac{a}{1+a}$$

because  $P(H_0|\vec{x}) = 1 - P(H_1|\vec{x})$ .

- (b) Reviewer B suggests that  $H_1$  should be  $\pi = 0.4$  instead. What is the posterior probability of  $H_0$  under this Bayesian comparison?

### Exercise 5.7

**Bayesian confidence intervals.** The `binom.bayes()` function in the `binom` package permits the calculation of Bayesian confidence intervals over  $\pi$  for various numbers of successes  $x$ , total trials  $n$ , and  $a$  and  $b$  (specified as `prior.shape1` and `prior.shape2` respectively—but `prior.shape1=a-1` and `prior.shape2=b-1`). Install the `binom` package with the command

```
install.packages("binom")
```

and then use the `binom.bayes()` function to plot the size of a 95% confidence interval on  $\pi$  as a function of the total number of trials  $n$ , ranging from 10 to 10000 (in multiples of 10), where 70% of the trials are always successes. (Hold  $a$  and  $b$  constant, as values of your choice.)

### Exercise 5.8

**Contextual dependency in phoneme sequences.**

1. Reproduce the Bayesian hypothesis test of Section 5.2.3 for uniform priors with the following sequences, computing  $P(H_1|\mathbf{y})$  for each sequence. One of the three sequences was generated from a context-independent distribution, whereas the other two were generated from context-dependent distributions. Which one is most strongly indicated by the Bayesian hypothesis test to be generated from the context-independent distribution?
  - (a) A B A B B B B B A B B B A A A B B B B B B
  - (b) B A B B A B A B A A B A B A B B A A B A B

(c) B B B B A A A A B B B B B A A A A B B B B

2. Although we put uniform priors on all success parameters in Section 5.2.3, in the contextual-dependence model it makes more sense to have a SPARSE prior—that is, one that favors strong preferences for some phonemes over others after each type of context. A sparse beta prior is one for which at least one  $\alpha_i$  parameter is low ( $< 1$ ). Revise the model so that the prior on  $\pi_\emptyset$  remains uniform, but that  $\pi_A$  and  $\pi_B$  have symmetric  $\langle \alpha, \alpha \rangle$  priors (give both  $\pi_A$  and  $\pi_B$  the same prior). Plot the posterior probabilities  $P(H_1|\mathbf{y})$  for sequences (i-iii) as a function of  $\alpha$  for  $0 < \alpha \leq 1$ . What is the value of  $\alpha$  for which the context-independent sequence is most strongly differentiated from the context-dependent sequences (i.e. the differences in  $P(H_1|\mathbf{y})$  between sequence pairs are greatest)?

### Exercise 5.9

#### Frequentist confidence intervals.

In this problem you'll be calculating some frequentist confidence intervals to get a firmer sense of exactly what they are.

1. The `english` dataset in `languageR` has lexical-decision and naming reaction times (`RTlexdec` and `RTnaming` respectively) for 2197 English words. Plot histograms of the mean RT of each item. Calculate 95% frequentist confidence intervals for lexical-decision and naming times respectively, as described in Lecture 7, Section 2. Which experimental method gives tighter confidence intervals on mean RT?
2. The `t.test()` function, when applied to a set of data, returns a list whose component `conf.int` is the upper and lower bounds of a 95% confidence interval:

```
> x <- rnorm(100, 2)
> t.test(x)$conf.int

[1] 1.841554 2.307926
attr(,"conf.level")
[1] 0.95
```

Show that the procedure used in Section 5.3 gives the same results as using `t.test()` for the confidence intervals for the English lexical decision and naming datasets.

3. Not all confidence intervals generated from an “experiment” are the same size. For “experiments” consisting of 10 observations drawn from a standard normal distribution, use R to calculate a histogram of lengths of 95% confidence intervals on the mean. What is the shape of the distribution of confidence interval lengths? For this problem, feel free to use `t.test()` to calculate confidence intervals.

## Exercise 5.10

### Comparing two samples.

In class we covered confidence intervals and the one-sample  $t$ -test. This approach allows us to test whether a dataset drawn from a(n approximately) normally distributed population departs significantly from has a particular mean. More frequently, however, we are interested in comparing two datasets  $\mathbf{x}$  and  $\mathbf{y}$  of sizes  $n_x$  and  $n_y$  respectively, and inferring whether or not they are drawn from the same population. For this purpose, the TWO-SAMPLE  $t$ -TEST is appropriate.<sup>3</sup>

1. **Statistical power.** Suppose you have two populations and you can collect  $n$  total observations from the two populations. Intuitively, how should you distribute your observations among the two populations to achieve the greatest STATISTICAL POWER in a test that the two populations follow the same distribution?
2. Check your intuitions. Let  $n = 40$  and consider all possible values of  $n_x$  and  $n_y$  (note that  $n_y = n - n_x$ ). For each possible value, run 1000 experiments where the two populations are  $X \sim \mathcal{N}(0, 1)$  and  $Y \sim \mathcal{N}(1, 1)$ . Plot the power of the two-sample  $t$ -test at the  $\alpha = 0.05$  level as a function of  $n_x$ .
3. **Paired  $t$ -tests.** Sometimes a dataset can be naturally thought of as consisting of pairs of measurements. For example, if a phonetician measured voice onset time for the syllables [ba] and [bo] for many different speakers, the data could be grouped into a matrix of the form

|         | Syllable |          |
|---------|----------|----------|
| Speaker | [ba]     | [bo]     |
| 1       | $x_{11}$ | $x_{12}$ |
| 2       | $x_{21}$ | $x_{22}$ |
|         | $\vdots$ |          |

If we wanted to test whether the voice onset times for [ba] and [bo] came from the same distribution, we could simply perform a two-sample  $t$ -test on the data in column 1 versus the data in column 2.

On the other hand, this doesn't take into account the systematic differences in voice-onset time that may hold across speakers. What we might really want to do is test whether the *differences* between  $x_{i1}$  and  $x_{i2}$  are clustered around zero—which would indicate that the two data vectors probably do come from the same population—or around some non-zero number. This comparison is called a PAIRED  $t$ -TEST.

---

<sup>3</sup>For completeness, the statistic that is  $t$ -distributed for the two-sample test is:

$$\frac{\bar{\mathbf{y}} - \bar{\mathbf{x}}}{\sqrt{\sigma^2 \left( \frac{1}{n_x} + \frac{1}{n_y} \right)}}$$

where  $\bar{\mathbf{x}}$  and  $\bar{\mathbf{y}}$  are the sample means; in general, the variance  $\sigma^2$  is unknown and is estimated as  $\hat{\sigma}^2 = \frac{\sum_i (x_i - \bar{\mathbf{x}})^2 + \sum_i (y_i - \bar{\mathbf{y}})^2}{N-2}$  where  $N = n_x + n_y$ .

The file `spillover_word_rts` contains the average reading time (in milliseconds) of the second “spillover” word after a critical manipulation in self-paced reading experiment, for 52 sentence pairs of the form:

The children went outside to **play** early in the afternoon. (Expected)  
The children went outside to **chat** early in the afternoon. (Unexpected)

In a separate sentence completion study, 90% of participants completed the sentence

The children went outside to —

with the word *play*, making this the Expected condition. In these examples, the word whose reading time (RT) is measured would be *in*, as it appears two words after the critical word (in bold).

- (a) Use paired and unpaired *t*-tests to test the hypothesis that mean reading times at the second spillover word differ significantly in the Expected and Unexpected conditions. Which test leads to a higher significance value?
- (b) Calculate the correlation between the RTs for the unexpected and expected conditions of each item. Intuitively, should higher correlations lead to an increase or drop in statistical power for the paired test over the unpaired test? Why?

### Exercise 5.11: Non-overlapping standard errors.

You and your colleague measure the F1 formant frequency in pronunciation of the vowel [a] for two groups of 50 native speakers of English, one measurement for each speaker. The means and standard errors of these measurements are shown in Figure 5.5. Your colleague states, “we can be fairly confident in inferring that the two groups have significantly different mean F1 formant frequencies. As a rule of thumb, when you have a reasonably large number of measurements in each group and the standard error bars between the two groups are non-overlapping, we can reject the null hypothesis that the group means are the same at the  $p < 0.05$  level.” Is your colleague’s rule of thumb correct?

### Exercise 5.12: Log odds ratio versus correlation.

Are (log) odds ratios any different from correlation coefficients? Plot the relationship between log odds ratio and correlation coefficient for a number of different  $2 \times 2$  contingency tables. (If you want to take a sampling-based approach to exploring the space of possible contingency tables, you might use the Dirichlet distribution—see Section B.8—to randomly generate sets of cell probabilities).

### Exercise 5.13: Contingency tables.

1. The `dative` dataset in `languageR` contains data on sentences from the parsed Switchboard corpus (Godfrey et al., 1992) involving the dative-shift alternation, such as

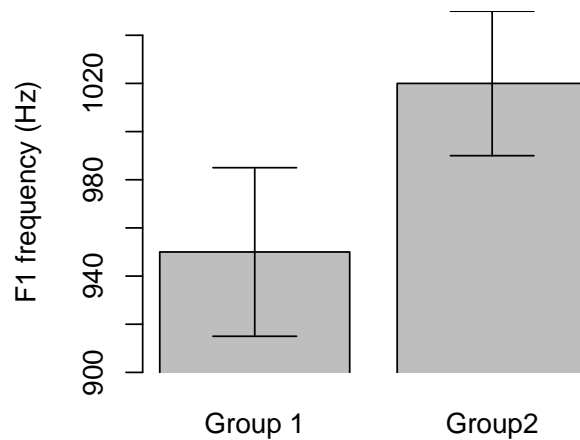


Figure 5.5: A bar plot with non-overlapping standard error bars

The actress gave **the toys** *to the children*. (Prepositional Object, PO)  
 The actress gave *the children* **the toys**. (Double Object, DO)

Irrespective of which alternate was used, it turns out that there are correlations among the properties of the theme (**the toys**) and the recipient (*the children*). See (Bresnan et al., 2007) for more details on the dataset and the coding of variables.

Definiteness and animacy are often found to be correlated. Look at the relationship between animacy and definiteness of (1) the theme, and (2) the recipient within the **dative** dataset, by constructing contingency tables and calculating the odds ratios in each case. (You can use `xtabs()` to construct contingency tables; look up the function in the index of Baayen.) For which semantic role are definiteness and animacy more strongly associated? Why do you think this might be the case? (Note that organizations, animals, intelligent machines, and vehicles were considered animate for this coding scheme (Zaenen et al., 2004)).

The `warlpiri` dataset in R's `languageR` package is a study of ergative case marking in elicited story descriptions in Warlpiri, an Australian Aboriginal language (O'Shannessy, 2006). Warlpiri is characterized by extremely free word order and a great deal of case marking. Below, for example, the ergative case marking (ERG) on the first word of the sentence identifies it as the subject of the sentence:



Ngarrka- ngku ka wawirri panti- rni. (Hale, 1983)  
man ERG AUX kangaroo spear NONPAST

“The man is spearing the kangaroo”.

In some dialects of Warlpiri, however, using the ergative case is not obligatory. Note that there would be a semantic ambiguity if the case marking were eliminated from the first word, because neither “man” nor “kangaroo” would have case marking to indicate its semantic relation to the verb “spear”.

- (a) Which of the following variables have a significant association with ergative case marking?

animacy of subject  
word order

- (b) Which of the following variables have an effect on whether subject animacy and word order have a significant association with use of ergative case marking? (For each of the below variables, split the `warlpiri` dataset in two and do a statistical test of association on each half.)

overtness of object  
age group

```
> library(plyr)
> options(lineWidth = 60)
```

# Chapter 6

## Generalized Linear Models

In Chapters 2 and 4 we studied how to estimate simple probability densities over a single random variable—that is, densities of the form  $P(Y)$ . In this chapter we move on to the problem of estimating *conditional* densities—that is, densities of the form  $P(Y|X)$ . Logically speaking, it would be possible to deal with this problem simply by assuming that  $Y$  may have an arbitrarily different distribution for each possible value of  $X$ , and to use techniques we’ve covered already to estimate a different density  $P(Y|X = x_i)$  for each possible value  $x_i$  of  $X$ . However, this approach would miss the fact that  $X$  may have a *systematic* effect on  $Y$ ; missing this fact when it is true would make it much more difficult to estimate the conditional distribution. Here, we cover a popular family of conditional probability distributions known as GENERALIZED LINEAR MODELS. These models can be used for a wide range of data types and have attractive computational properties.

### 6.1 The form of the generalized linear model

Suppose we are interested in modeling the distribution of  $Y$  conditional on a number of random variables  $X_1, \dots, X_n$ . Generalized linear models are a framework for modeling this type of conditional distribution  $P(Y|X_1, \dots, X_n)$  subject to four key assumptions:

1. The influences of the  $\{X_i\}$  variables on  $Y$  can be summarized into an intermediate form, the LINEAR PREDICTOR  $\eta$ ;
2.  $\eta$  is a linear combination of the  $\{X_i\}$ ;
3. There is a smooth, invertible function  $l$  mapping  $\eta$  to the expected value  $\mu$  of  $Y$ ;
4. The distribution  $P(Y = y; \mu)$  of  $Y$  around  $\mu$  is a member of a certain class of noise functions and is not otherwise sensitive to the  $X_i$  variables.<sup>1</sup>

Assumptions 1 through 3 can be expressed by the following two equations:

---

<sup>1</sup>The class of allowable noise functions is described in Section ??.

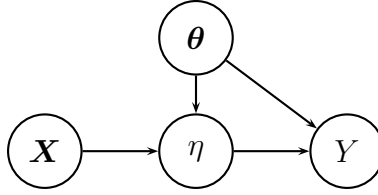


Figure 6.1: A graphical depiction of the generalized linear model. The influence of the conditioning variables  $\mathbf{X}$  on the response  $Y$  is completely mediated by the linear predictor  $\eta$ .

$$\eta = \alpha + \beta_1 X_1 + \cdots + \beta_n X_n \quad \text{(linear predictor)} \quad (6.1)$$

$$\eta = l(\mu) \quad \text{(link function)} \quad (6.2)$$

Assumption 4 implies conditional independence of  $Y$  from the  $\{X_i\}$  variables given  $\eta$ .

Various choices of  $l(\mu)$  and  $P(Y = y; \mu)$  give us different classes of models appropriate for different types of data. In all cases, we can estimate the parameters of the models using any of likelihood-based techniques discussed in Chapter 4. We cover three common classes of such models in this chapter: linear models, logit (logistic) models, and log-linear models.

## 6.2 Linear models and linear regression

We can obtain the classic LINEAR MODEL by choosing the identity link function

$$\eta = l(\mu) = \mu$$

and a noise function that adds noise

$$\epsilon \sim N(0, \sigma^2)$$

to the mean  $\mu$ . Substituting these in to Equations (6.1) and 6.2, we can write  $Y$  directly as a function of  $\{X_i\}$  as follows:

$$Y = \underbrace{\alpha + \beta_1 X_1 + \cdots + \beta_n X_n}_{\text{Predicted Mean}} + \underbrace{\epsilon}_{\text{Noise} \sim N(0, \sigma^2)} \quad (6.3)$$

We can also write this whole thing in more compressed form as  $Y \sim N(\alpha \sum_i \beta_i X_i, \sigma^2)$ .

To gain intuitions for how this model places a conditional probability density on  $Y$ , we can visualize this probability density for a single independent variable  $X$ , as in Figure 6.2—lighter means more probable. Each vertical slice of constant  $X = x$  represents a conditional

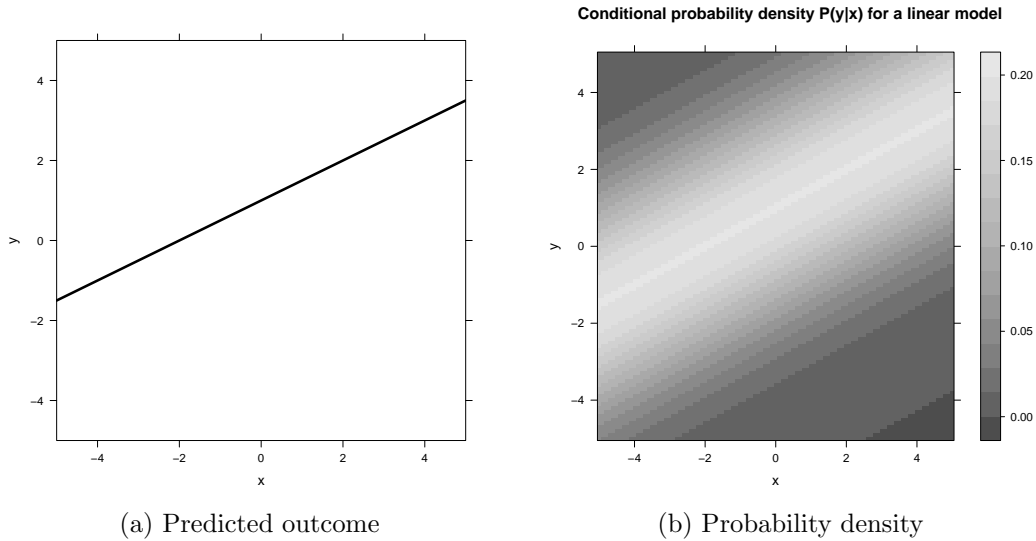


Figure 6.2: A plot of the probability density on the outcome of the  $Y$  random variable given the  $X$  random variable; in this case we have  $\eta = \frac{1}{2}X$  and  $\sigma^2 = 4$ .

distribution  $P(Y|x)$ . If you imagine a vertical line extending through the plot at  $X = 0$ , you will see that the plot along this line is lightest in color at  $Y = -1 = \alpha$ . This is the point at which  $\epsilon$  takes its most probable value, 0. For this reason,  $\alpha$  is also called the **INTERCEPT** parameter of the model, and the  $\beta_i$  are called the **SLOPE** parameters.

### 6.2.1 Fitting a linear model

The process of estimating the parameters  $\alpha$  and  $\beta_i$  of a linear model on the basis of some data (also called **FITTING** the model to the data) is called **LINEAR REGRESSION**. There are many techniques for parameter estimation in linear regression; here we will cover the method of maximum likelihood and also Bayesian linear regression.

#### Maximum-likelihood linear regression

Before we talk about exactly what the maximum-likelihood estimate looks like, we'll introduce some useful terminology. Suppose that we have chosen model parameters  $\hat{\alpha}$  and  $\{\hat{\beta}_i\}$ . This means that for each point  $\langle x_j, y_j \rangle$  in our dataset  $\mathbf{y}$ , we can construct a **PREDICTED VALUE** for  $\hat{y}_j$  as follows:

$$\hat{y}_j = \hat{\alpha} + \hat{\beta}_1 x_{j1} + \dots + \hat{\beta}_n x_{jn}$$

where  $x_{ji}$  is the value of the  $i$ -th predictor variable for the  $j$ -th data point. This predicted value is both the expected value and the modal value of  $Y_j$  due to the Gaussian-noise assumption of linear regression. We define the **RESIDUAL** of the  $j$ -th data point simply as

$$y_j - \hat{y}_j$$

—that is, the amount by which our model’s prediction missed the observed value.

It turns out that for linear models with a normally-distributed error term  $\epsilon$ , the log-likelihood of the model parameters with respect to  $\mathbf{y}$  is proportional to the sum of the squared residuals. This means that the maximum-likelihood estimate of the parameters is also the estimate that minimizes the the sum of the squared residuals. You will often see description of regression models being fit using LEAST-SQUARES estimation. Whenever you see this, recognize that this is equivalent to maximum-likelihood estimation under the assumption that residual error is normally-distributed.

### 6.2.2 Fitting a linear model: case study

The dataset `english` contains reaction times for lexical decision and naming of isolated English words, as well as written frequencies for those words. Reaction times are measured in milliseconds, and word frequencies are measured in appearances in a 17.9-million word written corpus. (All these variables are recorded in log-space) It is well-established that words of high textual frequency are generally responded to more quickly than words of low textual frequency. Let us consider a linear model in which reaction time  $RT$  depends on the log-frequency,  $F$ , of the word:

$$RT = \alpha + \beta_F F + \epsilon \tag{6.4}$$

This linear model corresponds to a FORMULA in R, which can be specified in either of the following ways:

`RT ~ F`  
`RT ~ 1 + F`

An  
Intro-  
duction  
to R,  
section  
11.1

The `1` in the latter formula refers to the intercept of the model; the presence of an intercept is implicit in the first formula.

The result of the linear regression is an intercept  $\alpha = 843.58$  and a slope  $\beta_F = -29.76$ . The `WrittenFrequency` variable is in natural log-space, so the slope can be interpreted as saying that if two words differ in frequency by a factor of  $e \approx 2.718$ , then on average the more frequent word will be recognized as a word of English 26.97 milliseconds faster than the less frequent word. The intercept, 843.58, is the predicted reaction time for a word whose log-frequency is 0—that is, a word occurring only once in the corpus.

### 6.2.3 Conceptual underpinnings of best linear fit

Let us now break down how the model goes about fitting data in a simple example.

Suppose we have only three observations of log-frequency/RT pairs:

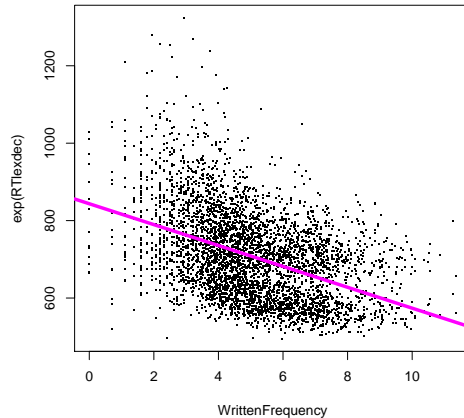


Figure 6.3: Lexical decision reaction times as a function of word frequency

$$\begin{aligned} &\langle 4, 800 \rangle \\ &\langle 6, 775 \rangle \\ &\langle 8, 700 \rangle \end{aligned}$$

Let us consider four possible parameter estimates for these data points. Three estimates will draw a line through two of the points and miss the third; the last estimate will draw a line that misses but is reasonably close to all the points.

First consider the solid black line, which has intercept 910 and slope -25. It predicts the following values, missing all three points:

| $x$ | $\hat{y}$ | Residual ( $\hat{y} - y$ ) |
|-----|-----------|----------------------------|
| 4   | 810       | -10                        |
| 6   | 760       | 15                         |
| 8   | 710       | -10                        |

and the sum of its squared residuals is 425. Each of the other three lines has only one non-zero residual, but that residual is much larger, and in all three cases, the sum of squared residuals is larger than for the solid black line. This means that the likelihood of the parameter values  $\alpha = 910, \beta_F = -25$  is higher than the likelihood of the parameters corresponding to any of the other lines.

What is the MLE for  $\alpha, \beta_F$  with respect to these three data points, and what are the residuals for the MLE?

Results of a linear fit (almost every statistical software package supports linear regression) indicate that the MLE is  $\alpha = 908\frac{1}{3}, \beta = -25$ . Thus the MLE has the same slope as the solid

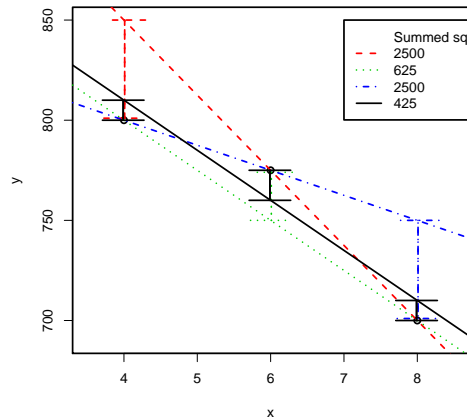


Figure 6.4: Linear regression with three points

black line in Figure 6.4, but the intercept is slightly lower. The sum of squared residuals is slightly better too.

**Take-home point:** for linear regression, getting everything wrong by a little bit is better than getting a few things wrong by a lot.

## 6.3 Handling multiple predictors

In many cases, we are interested in simultaneously investigating the linear influence of two or more predictor variables on a single response. We’ll discuss two methods of doing this: RESIDUALIZING and MULTIPLE LINEAR REGRESSION.

As a case study, consider naming reaction times from the `english` dataset, and now imagine that we’re interested in the influence of orthographic neighbors. (An orthographic neighbor of a word  $w$  is one that shares most of its letters with  $w$ ; for example, *cat* has several orthographic neighbors including *mat* and *rat*.) The `english` dataset summarizes this information in the `Ncount` variable, which measures ORTHOGRAPHIC NEIGHBORHOOD DENSITY as (I believe) the number of maximally close orthographic neighbors that the word has. How can we investigate the role of orthographic neighborhood while simultaneously taking into account the role of word frequency?

### 6.3.1 Residualizing

One approach would be a two-step process: first, construct a linear regression with frequency as the predictor and RT as the response. (This is commonly called “regressing RT against frequency”.) Second, construct a new linear regression with neighborhood density as the predictor the *residuals from the first regression* as the response. The transformation of a raw RT into the residual from a linear regression is called RESIDUALIZATION. Figure 6.5):

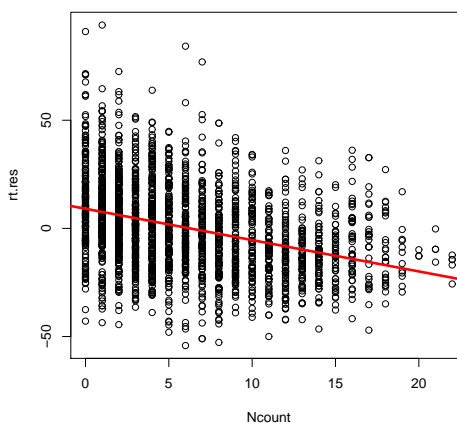


Figure 6.5: Plot of frequency-residualized word naming times and linear regression against neighborhood density

```
> english.young <- subset(english, AgeSubject == "young")
> attach(english.young)
> rt.freq.lm <- lm(exp(RTnaming) ~ WrittenFrequency)
> rt.freq.lm
```

Call:

```
lm(formula = exp(RTnaming) ~ WrittenFrequency)
```

Coefficients:

| (Intercept) | WrittenFrequency |
|-------------|------------------|
| 486.506     | -3.307           |

```
> rt.res <- resid(rt.freq.lm)
> rt.ncount.lm <- lm(rt.res ~ Ncount)
> plot(Ncount, rt.res)
> abline(rt.ncount.lm, col = 2, lwd = 3)
> detach()
> rt.ncount.lm
```

Call:

```
lm(formula = rt.res ~ Ncount)
```

Coefficients:

| (Intercept) | Ncount |
|-------------|--------|
| 9.080       | -1.449 |



Even after linear effects of frequency have been accounted for by removing them from the RT measure, neighborhood density still has some effect – words with higher neighborhood density are named more quickly.

### 6.3.2 Multiple linear regression

The alternative is to build a single linear model with more than one predictor. A linear model predicting naming reaction time on the basis of both frequency  $F$  and neighborhood density  $D$  would look like this:

$$RT = \alpha + \beta_F F + \beta_D D + \epsilon$$

and the corresponding R formula would be either of the following:

```
RT ~ F + D
RT ~ 1 + F + D
```

Plugging this in gives us the following results:

```
> rt.both.lm <- lm(exp(RTnaming) ~ WrittenFrequency + Ncount, data = english.young)
> rt.both.lm
```

Call:

```
lm(formula = exp(RTnaming) ~ WrittenFrequency + Ncount, data = english.young)
```

Coefficients:

| (Intercept) | WrittenFrequency | Ncount |
|-------------|------------------|--------|
| 493.638     | -2.899           | -1.465 |

Note that the results are qualitatively similar but quantitatively different than for the residualization approach: larger effect sizes have been estimated for both `WrittenFrequency` and `Ncount`.

## 6.4 Confidence intervals and hypothesis testing for linear regression

Just as there was a close connection between hypothesis testing with the one-sample  $t$ -test and a confidence interval for the mean of a sample, there is a close connection between hypothesis testing and confidence intervals for the parameters of a linear model. We'll start by explaining the confidence interval as the fundamental idea, and see how this leads to hypothesis tests.

Figure 6.6 illustrates the procedures by which confidence intervals are constructed for a sample mean (one parameter) and for the intercept and slope of a linear regression with one

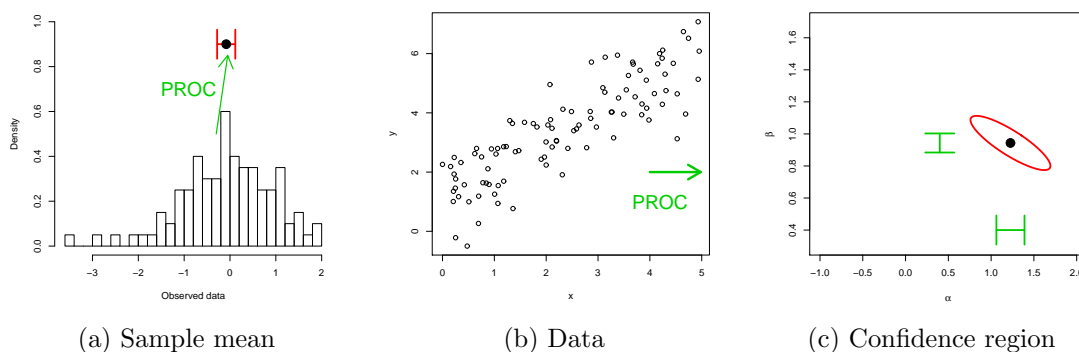


Figure 6.6: The confidence-region construction procedure for (a) sample means and (b-c) parameters of a linear model. The black dots are the maximum-likelihood estimates, around which the confidence regions are centered.

predictor. In both cases, a dataset  $\mathbf{y}$  is obtained, and a fixed procedure is used to construct boundaries of a CONFIDENCE REGION from  $\mathbf{y}$ . In the case of the sample mean, the “region” is in one-dimensional space so it is an interval. In the case of a linear regression model, the region is in two-dimensional space, and looks like an ellipse. The size and shape of the ellipse are determined by the VARIANCE-COVARIANCE MATRIX of the linear predictors, and are determined using the fact that the joint distribution of the estimated model parameters is multivariate-normal distributed (Section 3.5). If we collapse the ellipse down to only one dimension (corresponding to one of the linear model’s parameters), we have a confidence interval on that parameter; this one-dimensional confidence interval is  $t$  distributed with  $N - k$  degrees of freedom (Section B.5), where  $N$  is the number of observations and  $k$  is the number of parameters in the linear model.<sup>2</sup>

We illustrate this in Figure 6.7 for the linear regression model of frequency against word naming latency. The model is quite certain about the parameter estimates; however, note that there is a correlation between the parameter estimates. According to the analysis, if we reran this regression many times by repeatedly drawing data from the same population and estimating parameters, whenever the resulting intercept (i.e., average predicted RT for the rarest class of word) is higher, the facilitative effect of written frequency would tend to be larger, and vice versa. This is intuitively sensible because the most important thing for the regression is where it passes through the centroid of the data; so that if the intercept drops a bit, the slope has to rise to compensate.

Perhaps a more interesting example is looking at the confidence region obtained for the parameters of two predictors. In the literature on word recognition, for example, there has been some discussion over whether word frequency or word familiarity drives variability in average word-recognition time (or whether both have independent effects). Because subjective ratings of word familiarity are strongly correlated with word frequency, it is em-

<sup>2</sup>Formally this corresponds to marginalizing over the estimates of the other parameters that you’re collapsing over.

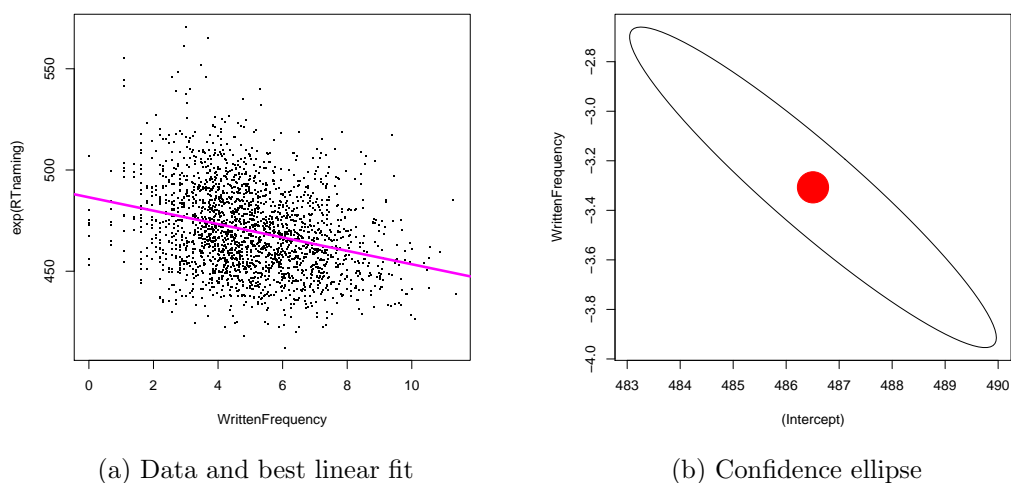


Figure 6.7: Confidence ellipse for parameters of regression of word naming latency against written frequency

pirically difficult to disentangle the two. Figure 6.8a shows a scatterplot of word familiarity against word frequency ratings for 2,197 English nouns and verbs (Spieler and Balota, 1997; Balota and Spieler, 1998); the empirical correlation is 0.79. The naming study carried out by Spieler and Balota (1997) was very large, and they obtained naming times for each of these words from 31 undergraduate native English speakers. A multiple linear regression analysis with frequency and familiarity as predictors puts 95% confidence intervals for their slopes in the linear model at  $[-2.49, -1.02]$  and  $[-4.33, -1.96]$  respectively. Hence we can conclude that each of frequency and familiarity contribute independently in determining naming time (insofar as the measurements of frequency and familiarity themselves are accurately measured).

However, this was a very large study, and one might reasonably ask what conclusions one could draw from a much smaller study. The same multiple linear regression based on a random subsample of 200 of the words from Spieler and Balota’s study gives us confidence intervals for the effects of word frequency and familiarity on naming time of  $[-3.67, 1.16]$  and  $[-7.26, 0.82]$ . With this smaller dataset, we cannot confidently conclude that either predictor is independently a determinant of naming time. Yet this negative result conceals an important conclusion that we can still draw. Figure 6.8 plots confidence *regions* for the two model parameters, as well as confidence intervals for each individual parameter, in models of the full dataset (solid green lines) and the reduced, 200-word dataset (dashed magenta lines). Although the reduced-dataset confidence region shows that we cannot be confident that either parameter is negative (i.e. that it has a facilitatory effect on naming time), we can be quite confident that it is not the case that *both* parameters are non-negative: the ellipse does not come close to encompassing the origin. That is, we can be confident that *some combination* of word frequency and familiarity has a reliable influence on naming time. We return to this point in Section 6.5.2 when we cover how to compare models differing by more

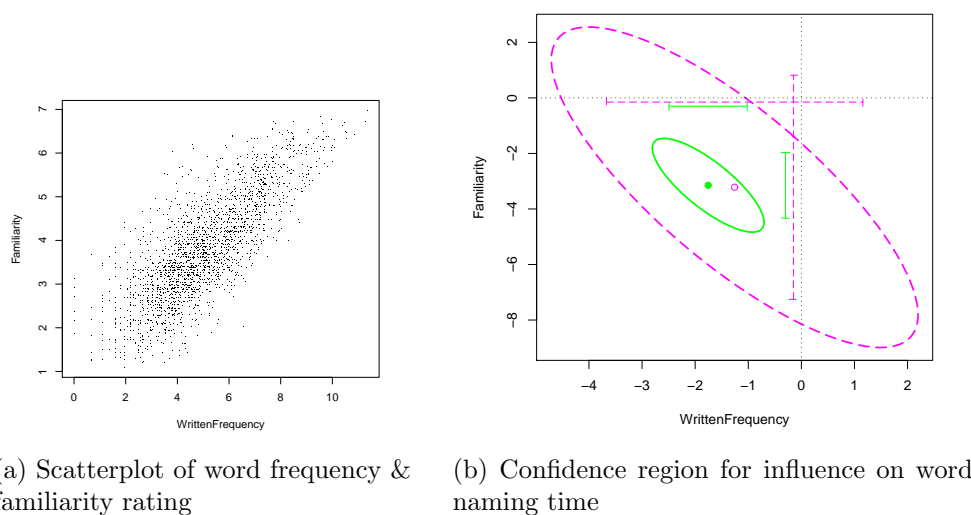


Figure 6.8: Confidence region on written word frequency and word familiarity for full dataset of Spieler and Balota (1997), and reduced subset of 200 items

than one parameter through the  $F$  test.

## 6.5 Hypothesis testing in multiple linear regression

An extremely common use of linear models is in testing hypotheses regarding whether one or more predictor variables have a reliable influence on a continuously-distributed response. Examples of such use in the study of language might include but are not limited to:

- Does a word's frequency reliably influence how rapidly it is recognized, spoken, or read?
- Are words of different parts of speech recognized, spoken, or read at different rates above and beyond the effects of word frequency (and perhaps other properties such as word length)?
- Does the violation of a given syntactic constraint affect a native speaker's rating of the felicity of sentences with the violation (as compared to sentences without the violation)?
- Does the context in which a sound is uttered reliably influence one of its phonetic properties (such as voice-onset time for stops, or format frequency for vowels)?

All of these questions may be addressed within the Neyman-Pearson paradigm frequentist hypothesis-testing paradigm introduced in Section 5.4. Recall that the Neyman-Pearson paradigm involves specifying a null hypothesis  $H_0$  and determining whether to reject it in favor of a more general and complex hypothesis  $H_A$ . In many cases, we are interested in

comparing whether a more complex linear regression is justified by the data over a simpler regression. Under these circumstances, we can take the simpler model  $M_0$  as the null hypothesis, and the more complex model  $M_A$  as the alternative hypothesis. There are two statistical tests that you will generally encounter for this purpose: one based on the  $t$  statistic (which we already saw in Section 5.3) and another, the  $F$  test, which is based on what is called the  $F$  statistic. However, the former is effectively a special case of the latter, so here we'll look at how to use the  $F$  statistic for hypothesis testing with linear models; then we'll briefly cover the use of the  $t$  statistic for hypothesis testing as well.

### 6.5.1 Decomposition of variance

The  $F$  test takes advantage of a beautiful property of linear models to compare  $M_0$  and  $M_A$ : the DECOMPOSITION OF VARIANCE. Recall that the VARIANCE of a sample is simply the sum of the square deviations from the mean:

$$\text{Var}(\mathbf{y}) = \sum_j (y_j - \bar{y})^2 \quad (6.5)$$

where  $\bar{y}$  is the mean of the sample  $\mathbf{y}$ . For any model  $M$  that predicts values  $\hat{y}_j$  for the data, the RESIDUAL VARIANCE of  $M$  is quantified in exactly the same way:

$$\text{Var}_M(\mathbf{y}) = \sum_j (y_j - \hat{y}_j)^2 \quad (6.6)$$

A beautiful and extraordinarily useful property of linear models is that the sample variance can be split apart, or *partitioned*, into (a) the component that is explained by  $M$ , and (b) the component that remains unexplained by  $M$ . This can be written as follows (see Exercise 6.5):

$$\text{Var}(\mathbf{y}) = \overbrace{\sum_j (y_j - \hat{y}_j)^2}^{\text{Var}_M(\mathbf{y})} + \overbrace{\sum_j (\hat{y}_j - \bar{y})^2}^{\text{unexplained}} \quad (6.7)$$

Furthermore, if two models are nested (i.e., one is a special case of the other), then the variance can be further subdivided among those two models. Figure 6.9 shows the partitioning of variance for two nested models.

### 6.5.2 Comparing linear models: the $F$ test statistic

The  $F$  test is widely used for comparison of linear models, and forms the foundation for many analytic techniques including the analysis of variance (ANOVA).

Recall that our starting assumption in this section was that we had two linear models: a more general model  $M_A$ , and a special case  $M_0$ —that is, for any instance of  $M_0$  we can

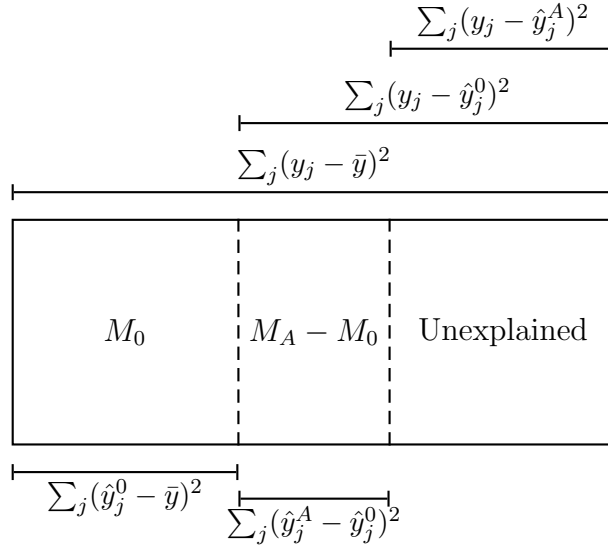


Figure 6.9: The partitioning of residual variance in linear models. Symbols in the box denote the variance explained by each model; the sums outside the box quantify the variance in each combination of sub-boxes.

achieve an instance of  $M_A$  with the same distribution over  $Y$  by setting the parameters appropriately. In this situation we say that  $M_0$  is NESTED inside  $M_A$ . Once we have found maximum-likelihood estimates of  $M_0$  and  $M_A$ , let us denote their predicted values for the  $j$ -th observation as  $\hat{y}_j^0$  and  $\hat{y}_j^A$  respectively. Then, if  $M_0$  has  $k_0$  parameters and  $M_A$  has  $k_A$  parameters, the  $F$  statistic, defined below, is widely used for a frequentist hypothesis test with  $M_0$  the null hypothesis and  $M_A$  the alternative hypothesis:

$$F = \frac{\sum_j (\hat{y}_j^A - \hat{y}_j^0)^2 / (k_A - k_0)}{\sum_j (y_j - \hat{y}_j^A)^2 / (n - k_A)} \quad (6.8)$$

Because of the decomposition of variance, the  $F$  statistic can also be written as follows:

$$F = \frac{\left[ \sum_j (y_j - \hat{y}_j^0)^2 - \sum_j (y_j - \hat{y}_j^A)^2 \right] / (k_A - k_0)}{\left[ \sum_j (y_j - \hat{y}_j^A)^2 \right] / (n - k_A)}$$

which underscores that the  $F$  statistic compares the amount of regularity explained by  $M_A$  unexplained by  $M_0$  (the numerator) with the amount of regularity unexplained by  $M_A$  (the denominator). The numerator and denominator are often called MEAN SQUARES.

Because of the decomposition of variance, the  $F$  test can be given a straightforward geometric interpretation. Take a look at the labels on the boxes in Figure 6.9 and convince yourself that the sums in the numerator and the denominator of the  $F$  statistic correspond

respectively to the boxes  $M_A - M_0$  and Unexplained. Thus, using the  $F$  statistic for hypothesis testing is often referred to as evaluation of the RATIO OF THE SUMS OF MEAN SQUARES.

For a comparison of models with  $k_0$  and  $k_A$  parameters on a dataset with  $n$  observations, the  $F$  statistic follows the  $F$  distribution (Section B.6)—which is parameterized by *two* degrees of freedom—with  $(k_A - k_0, n - k_A)$  degrees of freedom. Because of its importance for frequentist statistical inference in linear models, the  $F$  distribution has been worked out in detail and is accessible in most statistical software packages.

### 6.5.3 Model comparison: case study

We can bring the  $F$  test to bear in our investigation of the relative contributions of word frequency and familiarity on word naming latency; we will focus on analysis of the reduced 200-item dataset. First let us consider a test in which the null hypothesis  $H_0$  is that only word frequency has a reliable effect, and the alternative hypothesis  $H_A$  is that both word frequency (or “Freq” for short) and familiarity (“Fam”) have reliable effects.  $H_0$  corresponds to the model  $RT \mathcal{N}(\alpha + \beta_{\text{Freq}} \text{Freq}, \sigma^2)$ ;  $H_A$  corresponds to the model  $RT \mathcal{N}(\alpha + \beta_{\text{Freq}} \text{Freq} + \beta_{\text{Fam}} \text{Fam}, \sigma^2)$ . After obtaining maximum-likelihood fits of both models, we can compute the residual sums of squares  $\sum_j (\hat{y}_j - y_j)^2$  for each model; these turn out to be 76824.28 for  $M_0$  and 75871.58 for  $M_A$ .  $M_0$  has two parameters,  $M_A$  has three, and we have 200 observations; hence  $k_A - k_0 = 1$  and  $N - k_A = 197$ . The  $F$  statistic for our hypothesis test is thus

$$\begin{aligned} F &= \frac{[76824.28 - 75871.58] / 1}{[75871.58] / 197} \\ &= 2.47 \end{aligned}$$

with (1,197) degrees of freedom. Consulting the cumulative distribution function for the  $F$  statistic we obtain a  $p$ -value of 0.12.

We can also apply the  $F$  test for comparisons of models differing in multiple parameters, however. For example, let  $M'_0$  be a model in which neither word frequency nor familiarity has an effect on naming time. The residual variance in this model is the entire sample variance, or 82270.49. For a comparison between  $M'_0$  and  $M_A$  we obtain an  $F$  statistic of

$$\begin{aligned} F &= \frac{[82270.49 - 75871.58] / 2}{[75871.58] / 197} \\ &= 8.31 \end{aligned}$$

with (2,197) degrees of freedom. The corresponding  $p$ -value is 0.00034, indicating that our data are extremely unlikely under  $M'_0$  and that  $M_A$  is far preferable. Thus, although we could not adjudicate between word frequency and familiarity with this smaller dataset, we could say confidently that *some combination of the two* has a reliable effect on word naming time.

Another widely used test for the null hypothesis that within a  $k$ -parameter model, a *single* parameter  $\beta_i$  is 0. This hypothesis can be tested through a  $t$ -test where the  $t$  statistic is the ratio of the parameter estimate,  $\hat{\beta}_i$  to the standard error of the estimate,  $\text{SE}(\hat{\beta}_i)$ . For a dataset with  $N$  observations, this  $t$ -test has  $N - k$  degrees of freedom. However, a  $F$  statistic with  $(1, m)$  has the same distribution as the square of a  $t$  statistic with  $m$  degrees of freedom. For this reason, the  $t$ -test for linear models can be seen as a special case of the more general  $F$  test; the latter can be applied to compare nested linear models differing in any number of parameters.

## 6.6 Analysis of Variance

Recall that we just covered linear models, which are conditional probability distributions of the form

$$P(Y|X) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon \quad (6.9)$$

where  $\epsilon \sim \mathcal{N}(0, \sigma^2)$ . We saw how this paradigm can be put to use for modeling the predictive relationship of continuous variables, such as word frequency, familiarity, and neighborhood density, on reaction times in word recognition experiments.

In many cases, however, the predictors of interest are not continuous. For example, for the `english` dataset in `languageR` we might be interested in how naming times are influenced by the type of the initial phoneme of the word. This information is coded by the `Frication` variable of the dataset, and has the following categories:

|                        |  |
|------------------------|--|
| <code>burst</code>     | the word starts with a burst consonant     |
| <code>frication</code> | the word starts with a fricative consonant |
| <code>long</code>      | the word starts with a long vowel          |
| <code>short</code>     | the word starts with a short vowel         |

It is not obvious how these categories might be meaningfully arranged on the real number line. Rather, we would simply like to investigate the possibility that the mean naming time differs as a function of initial phoneme type.

The most widespread technique used to investigate this type of question is the ANALYSIS OF VARIANCE (often abbreviated ANOVA). Although many books go into painstaking detail covering different instances of ANOVA, you can gain a firm foundational understanding of the core part of the method by thinking of it as a special case of multiple linear regression.

### 6.6.1 Dummy variables

Let us take the example above, where `Frication` is a categorical predictor. Categorical predictors are often called `FACTORS`, and the values they take are often called `LEVELS`. (This is also the nomenclature used in `R`.) In order to allow for the possibility that each level of the



factor could have arbitrarily different effects on mean naming latency, we can create DUMMY PREDICTOR VARIABLES, one per level of the factor:

| Level of Frication | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
|--------------------|-------|-------|-------|-------|
| <b>burst</b>       | 1     | 0     | 0     | 0     |
| <b>frication</b>   | 0     | 1     | 0     | 0     |
| <b>long</b>        | 0     | 0     | 1     | 0     |
| <b>short</b>       | 0     | 0     | 0     | 1     |

(Variables such as these which are 0 unless a special condition holds, in which case they are 1, are often referred to as INDICATOR VARIABLES). We then construct a standard linear model with predictors  $X_1$  through  $X_4$ :

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon \quad (6.10)$$

When we combine the dummy predictor variables with the linear model in (6.9), we get the following equations for each level of **Frication**:

| Level of Frication | Linear model                      |
|--------------------|-----------------------------------|
| <b>burst</b>       | $Y = \alpha + \beta_1 + \epsilon$ |
| <b>frication</b>   | $Y = \alpha + \beta_2 + \epsilon$ |
| <b>long</b>        | $Y = \alpha + \beta_3 + \epsilon$ |
| <b>short</b>       | $Y = \alpha + \beta_4 + \epsilon$ |

This linear model thus allows us to code a different predicted mean (and most-likely predicted value) for each level of the predictor, by choosing different values of  $\alpha$  and  $\beta_i$ .

However, it should be clear from the table above that only four distinct means can be predicted in this linear model—one for each level of **Frication**. We don't need five parameters (one for  $\alpha$  and four for the  $\beta_i$ ) to encode four means; one of the parameters is redundant. This is problematic when fitting the model because it means that there is no unique maximum-likelihood estimate.<sup>3</sup> To eliminate this redundancy, we arbitrarily choose one level of the factor as the **BASELINE** level, and we don't introduce a dummy predictor for the baseline level. If we choose **burst** as the baseline level,<sup>4</sup> then we can eliminate  $X_4$ , and make  $X_1, X_2, X_3$  dummy indicator variables for **frication**, **long**, and **short** respectively, giving us the linear model

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon \quad (6.11)$$

---

<sup>3</sup>For example, if  $\alpha = 0, \beta_1 = \beta_2 = \beta_3 = \beta_4 = 1$  is a maximum-likelihood estimate, then  $\alpha = 1, \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$  is as well because it encodes exactly the same model.

<sup>4</sup>By default, **R** chooses the first level of a factor as the baseline, and the first level of a factor is whatever level comes first alphabetically unless you specified otherwise when the factor was constructed—see the **levels** argument of the function **factor()** in the documentation.

where predicted means for the four classes are as follows:<sup>5</sup>

| Level of Frication | Predicted mean     |
|--------------------|--------------------|
| burst              | $\alpha$           |
| frication          | $\alpha + \beta_1$ |
| long               | $\alpha + \beta_2$ |
| short              | $\alpha + \beta_3$ |

### 6.6.2 Analysis of variance as model comparison

Now that we have completed the discussion of using dummy variables to construct a linear model with categorical predictors (i.e., factors), we shall move on to discussing what analysis of variance actually *does*. Consider that we now have two possible models of how word-initial frication affects naming time. We have the model of Equation (6.11) above, in which each class of frication predicts a different mean naming time, with noise around the mean distributed the same way for each class. We might also consider a simpler model in which frication has no effect on naming time. Such a model looks as follows:

$$Y = \alpha + \epsilon \tag{6.12}$$

Now look again at Figure 6.9 and think of the simpler model of Equation (6.12) as  $M_0$ , and the more complex model of Equation (6.11) as  $M_A$ . (Actually, the  $M_0$  explains *no* variance in this case because it just encodes the mean.) Because ANOVA is just a comparison of linear models, we can perform a hypothesis test between  $M_0$  and  $M_A$  by constructing an  $F$  statistic from the ratio of the amount of variance contained in the boxes  $M_A - M_0$  and Unexplained. The simpler model has one parameter and the more complex model has four, so we use Equation (6.8) with  $k_0 = 1, k_A = 4$  to construct the  $F$  statistic. The MLE of the single parameter for  $M_0$  (aside from the residual noise variance) is the sample mean  $\hat{\alpha} = 470$ , and the sum of squared residuals in this model is 1032186. For  $M_A$  with the dummy variable coding we've used, the MLEs are  $\hat{\alpha} = 471, \hat{\beta}_1 = 6, \hat{\beta}_2 = -4,$  and  $\hat{\beta}_3 = -16$ ; the sum of squared residuals is 872627. Thus the  $F$  statistic for this model comparison is

$$\begin{aligned} F(3, 2280) &= \frac{(1032186 - 872627)/3}{872627/2280} \\ &= 138.97 \end{aligned}$$

This  $F$  statistic corresponds to a  $p$ -value of  $1.089268e-82$ , yielding exceedingly clear evidence that the type of initial segment in a word affects its average naming latency.

---

<sup>5</sup>This choice of coding for the dummy variables is technically known as the choice of CONTRAST MATRIX. The choice of contrast matrix described here is referred to as the TREATMENT contrast matrix, or `contr.treatment` in R.

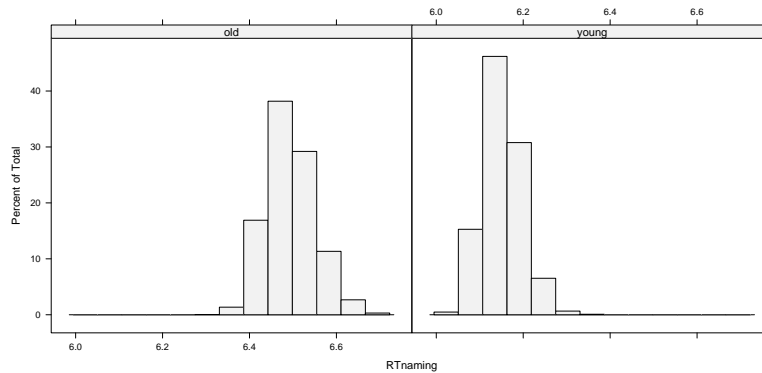


Figure 6.10: Histogram of naming latencies for young (ages  $\sim 22.6$ ) versus old (ages  $> 60$  speakers)

### 6.6.3 Testing for interactions

The `english` dataset includes average naming latencies not only for college-age speakers but also for speakers age 60 and over. This degree of age difference turns out to have a huge effect on naming latency (Figure 6.10):

```
histogram(~ RTnaming | AgeSubject, english)
```

Clearly, college-age speakers are faster at naming words than speakers over age 60. We may be interested in including this information in our model. In Lecture 10 we already saw how to include both variables in a multiple regression model. Here we will investigate an additional possibility: that different levels of frication may have different effects on mean naming latency depending on speaker age. For example, we might think that fricatives, which our linear model above indicates are the hardest class of word onsets, might be even harder for elderly speakers than they are for the young. When these types of inter-predictor contingencies are included in a statistical model they are called **INTERACTIONS**.

It is instructive to look explicitly at the linear model that results from introducing interactions between multiple categorical predictors. We will take `old` as the baseline value of speaker age, and leave `burst` as the baseline value of frication. This means that the “baseline” predictor set involves an old-group speaker naming a burst-initial word, and the intercept  $\alpha$  will express the predicted mean latency for this combination. There are seven other logically possible combinations of age and frication; thus our full model will have to have seven dummy indicator variables, each with its own parameter. There are many ways to set up these dummy variables; we’ll cover perhaps the most straightforward way. In addition to  $X_{\{1,2,3\}}$  for the non-baseline levels of frication, we add a new variable  $X_4$  for the non-baseline levels of speaker age (`young`). This set of dummy variables allows us to encode all eight possible groups, but it doesn’t allow us to estimate separate parameters for all these groups. To do this, we need to add three more dummy variables, one for each of the non-baseline frication levels when coupled with the non-baseline age level. This gives us the following complete set of codings:

| Frication | Age   | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|-----------|-------|-------|-------|-------|-------|-------|-------|-------|
| burst     | old   | 0     | 0     | 0     | 0     | 0     | 0     | 0     |
| frication | old   | 1     | 0     | 0     | 0     | 0     | 0     | 0     |
| long      | old   | 0     | 1     | 0     | 0     | 0     | 0     | 0     |
| short     | old   | 0     | 0     | 1     | 0     | 0     | 0     | 0     |
| burst     | young | 0     | 0     | 0     | 1     | 0     | 0     | 0     |
| frication | young | 1     | 0     | 0     | 1     | 1     | 0     | 0     |
| long      | young | 0     | 1     | 0     | 1     | 0     | 1     | 0     |
| short     | young | 0     | 0     | 1     | 1     | 0     | 0     | 1     |

We can test this full model against a strictly ADDITIVE model that allows for effects of both age and initial phoneme class, but not for interactions—that is, one with only  $X_{\{1,2,3,4\}}$ . It is critical to realize that the additive model is a *constrained* model: five parameters ( $\alpha$  and  $\beta_1$  through  $\beta_4$ ) cannot be used to encode eight arbitrary condition means within the linear framework. The best this  $M_0$  can do—the predicted condition means in its MLE—are compared with the true condition means below:

|             | Predicted in $M_0$ |        |             | Actual (and predicted in $M_A$ ) |        |
|-------------|--------------------|--------|-------------|----------------------------------|--------|
|             | Young              | Older  |             | Young                            | Older  |
| Burst       | 662.23             | 470.23 | Burst       | 661.27                           | 471.19 |
| Fricative   | 670.62             | 478.61 | Fricative   | 671.76                           | 477.48 |
| Long vowel  | 653.09             | 461.09 | Long vowel  | 647.25                           | 466.93 |
| Short vowel | 647.32             | 455.32 | Short vowel | 647.72                           | 454.92 |

The predicted per-category means in  $M_0$  can be recovered from the MLE parameter estimates:

$$\hat{\alpha} = 662.23 \quad \hat{\beta}_1 = 8.39 \quad \hat{\beta}_2 = -9.14 \quad \hat{\beta}_3 = -14.91 \quad \hat{\beta}_4 = -192$$

Recovering the predicted means in  $M_0$  from these parameter estimates is left as an exercise for the reader.

When the MLEs of  $M_0$  and  $M_A$  are compared using the  $F$ -test, we find that our  $F$ -statistic turns out to be  $F(3, 4560) = 2.69$ , or  $p = 0.0449$ . Hence we also have some evidence that initial segment type has different effects on average naming time dependign

#### 6.6.4 Repeated Measures ANOVA and Error Stratification

In the foregoing sections we have covered situations where all of the systematicity across observations can be summarized as deriving from predictors whose effects on the response are systematic and deterministic; all stochastic, idiosyncratic effects have been assumed to occur on level of the individual measurement of the response. In our analysis of average response times for recognition of English words, for example, we considered systematic effects of word frequency, familiarity, neighborhood density, and (in the case of word naming times) initial segment.

Yet it is a rare case in the study of language when there are no potential idiosyncratic effects that are incidental to the true interest of the researcher, yet affect entire *groups* of observations, rather than individual observations. As an example, Alexopoulou and Keller (2007) elicited quantitative subjective ratings of sentence acceptability in a study of pronoun resumption, embedding depth, and syntactic islands. One part of one of their experiments involved investigating whether there might be an interaction between embedding and the presence of a resumptive pronoun on sentence acceptability even in cases which are not syntactic islands (Ross, 1967). That is, among the four syntactic frames below, (1-b) should be much less acceptable than (1-a), but (1-d) should not be so much less acceptable than (1-c).

- |     |    |  |                           |
|-----|----|--|---------------------------|
| (1) | a. | Who will we fire —?                      | [UNEMBEDDED, −RESUMPTION] |
|     | b. | Who will we evict him?                   | [UNEMBEDDED, +RESUMPTION] |
|     | c. | Who does Lucy claim we will punish —?    | [EMBEDDED, −RESUMPTION]   |
|     | d. | Who does Emily claim we will arrest him? | [EMBEDDED, +RESUMPTION]   |

As is no doubt evident to the reader, even if we were to find that such a pattern holds for average acceptability ratings of these four sentences, a skeptic could reasonably object that the pattern might well result from the choice of words—the LEXICALIZATIONS—used to fill in the four syntactic templates. For example, *evict* is the least frequent of the four critical verbs above, and it is reasonable to imagine that sentences with less frequent words might tend to be rated as less acceptable.

Hence we want to ensure that our results *generalize* across the specific choice of lexicalizations used in this particular set of four sentences. One way of achieving this would be to prepare  $k > 1$  instances of syntactic frame, choosing a separate lexicalization randomly for each of the  $k$  instances of each frame ( $4k$  lexicalizations total). We might reasonably assume that the effects of choice of lexicalization on acceptability are normally distributed. Following our previous examples, we could use the following dummy-variable encodings:

|                           | $X_1$ | $X_2$ | $X_3$ |
|---------------------------|-------|-------|-------|
| [UNEMBEDDED, −RESUMPTION] | 0     | 0     | 0     |
| [UNEMBEDDED, +RESUMPTION] | 1     | 0     | 0     |
| [EMBEDDED, −RESUMPTION]   | 0     | 1     | 0     |
| [EMBEDDED, +RESUMPTION]   | 1     | 1     | 1     |

If  $\epsilon_L$  is the stochastic effect of the choice of lexicalization and  $\epsilon_E$  is the normally-distributed error associated with measuring the acceptability of a lexicalized frame, we get the following linear model:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon_L + \epsilon_E$$

Typically, we can think of speaker-level stochastic effects and measurement-level stochastic effects as independent; hence, because the sum of two independent normal random variables is itself normally distributed, we can just combine these two stochastic components of this equation:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

so we have a completely standard linear model. We could conduct hypothesis tests for this model in the same way as we have done previously in this chapter. For example, we could test the significance of an interaction between embedding and resumption—formally a comparison between a null-hypothesis model  $M_0$  in which  $\beta_3 = 0$  and an alternative model  $M_A$  with unconstrained  $\beta_3$ —by partitioning variance as in Table 6.1 and conducting an  $F$  test comparing the variance explained by adding  $\beta_3$  to the model with the residual variance left unexplained by  $M_A$ .

By choosing a different set of lexicalizations for each syntactic frame, however, we have introduced additional noise into our measurements that will only increase the difficulty of drawing reliable inferences regarding the effects of embedding, resumption, and their potential interaction. It turns out that we can in general do much better, by using the *same* lexicalizations for each syntactic frame. This is in fact what Alexopolou and Keller did, contrasting a total of nine sentence cohorts of the following types:

- (2) a. (i) Who will we fire \_\_\_? [UNEMBEDDED, -RESUMPTION]  
 (ii) Who will we fire him? [UNEMBEDDED, +RESUMPTION]  
 (iii) Who does Mary claim we will fire \_\_\_? [EMBEDDED, -RESUMPTION]  
 (iv) Who does Mary claim we will fire him? [EMBEDDED, +RESUMPTION]  
 b. (i) Who will we evict \_\_\_? [UNEMBEDDED, -RESUMPTION]  
 (ii) Who will we evict him? [UNEMBEDDED, +RESUMPTION]  
 (iii) Who does Elizabeth claim we will evict \_\_\_? [EMBEDDED, -RESUMPTION]  
 (iv) Who does Elizabeth claim we will evict him? [EMBEDDED, +RESUMPTION]  
 c. ...

Each cohort corresponds to a single lexicalization; in experimental studies such as these the more generic term *ITEM* is often used instead of lexicalization. This experimental design is often called *WITHIN-ITEMS* because the manipulation of ultimate interest—the choice of syntactic frame, or the *CONDITION*—is conducted for each individual item. Analysis of within-items designs using ANOVA is one type of what is called a *REPEATED-MEASURES ANOVA*, so named because multiple measurements are made for each of the items. The set of observations obtained for a single item thus constitute a *CLUSTER* that we hypothesize may have idiosyncratic properties that systematically affect the response variable, and which need to be taken into account when we draw statistical inferences regarding the generative process which gave rise to our data. For this reason, repeated-measures ANOVA is an analytic technique for what are known as *HIERARCHICAL MODELS*. Hierarchical models are themselves

an extremely rich topic, and we take them up in Chapter 8 in full detail. There is also, however, a body of analytic techniques which uses the partitioning of variance and  $F$  tests to analyze certain classes of hierarchical models using repeated-measures ANOVA. Because these techniques are extremely widespread in many literatures in the study of language and because these techniques do not require the full toolset for dealing with hierarchical models in general, we cover the repeated-measures ANOVA here. The reader is strongly encouraged, however, to compare the repeated-measure ANOVA with the analytic techniques introduced in Chapter 8, which ultimately offer greater overall flexibility and depth of analysis.

### Simple random-intercepts repeated-measures ANOVA

Exactly how to conduct repeated-measures ANOVA depends on the precise nature of the idiosyncratic cluster-level properties assumed. In our current example, the simplest scenario would be if each item (lexicalization) contributed the same fixed amount to average perceived acceptability regardless of the condition (syntactic frame) in which the lexicalization appeared. If we call the contribution of item  $i$  to acceptability  $a_i$ , then our model becomes

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + a_i + \epsilon$$

We may consider the  $a_i$  themselves to be stochastic: most canonically, they may be normally distributed around 0 with some unknown variance. Happily, the stochasticity in this model does not affect how we go about assessing the systematic effects— $\beta_1$  through  $\beta_3$ —of ultimate interest to us. We can partition the variance exactly as before.

#### 6.6.5 Condition-specific random effects and error stratification

More generally, however, we might consider the possibility that idiosyncratic cluster-level properties themselves *interact* with the manipulations we intend to carry out. In our case of embedding and resumption, for example, it could be the case that some of the verbs we choose might be particularly unnatural embedded in a complement clause, particularly natural with an overt resumptive-pronoun object, and/or particularly sensitive to specific combinations of embedding and resumptivity. Such a more general model would thus say that

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + a_i + b_{i1} X_1 + b_{i2} X_2 + b_{i3} X_3 + \epsilon$$

where  $\langle a_i, b_{1i}, b_{2i}, b_{3i} \rangle$  are jointly multivariate-normal with mean zero and some unknown covariance structure.

With this richer structure of idiosyncratic cluster-level properties, it turns out that we *cannot* partition the variance as straightforwardly as depicted in Figure ?? and draw reliable inferences in hypothesis tests about  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$ . It is instructive to step through the

precise reason for this. Suppose that we were to test for the presence of an interaction between resumption and embedding—that is, to test the null hypothesis  $M_0 : \beta_3 = 0$  against the alternative, more general  $M_A$ . Even if  $M_0$  is correct, in general the fit of  $M_A$  will account for more variance than  $M_0$  simply because  $M_A$  is a more expressive model. As in all cases, the amount of variance that  $M_A$  fails to explain will depend on the amount of noise at the level of specific observations (the variance of  $\epsilon$ ). But if  $M_0$  is true, the variance explained by  $M_A$  beyond  $M_0$  will depend not only the amount of observation-level noise but also on the amount and nature of cluster-level noise—that is, the variance of  $b_{i3}$  and its correlation with  $a_i$ ,  $b_{i1}$ , and  $b_{i2}$ . Exercise 6.7 asks you to demonstrate this effect through simulations.

Fortunately, there *does* turn out to be a way to test hypotheses in the face of such a rich structure of (normally-distributed) cluster-level properties: the STRATIFICATION OF VARIANCE. XXXXX

As a first example, let us simply examine the simple effect of adding a level of embedding to object-extracted cases without resumptive pronouns: Example (1-c) versus (1-c). In these cases, according to our dummy variable scheme we have  $X_1 = X_3 = 0$ , giving us the simplified linear equation:

$$Y = \alpha + \beta_2 X_2 + a_i b_{i2} X_2 + \epsilon \quad (6.13)$$

Figure 6.11 demonstrates the stratification of variance. Although everything except the box labeled “Residual Error” is part of the complete model of Equation (6.13), our  $F$ -test for the presence of a significant effect of embedding will pit the variance explained by embedding against the variance explained by idiosyncratic subject sensitivities to embedding condition.

Here is code that demonstrates the execution of the repeated-measures ANOVA:

```
> set.seed(2)
> library(mvtnorm)
> n <- 20
> m <- 20
> beta <- c(0.6,0.2) ## beta[1] corresponds to the intercept; beta[2] corresponds to t
> Sigma.b <- matrix(c(0.3,0,0,0.3),2,2) ## in this case, condition-specific speaker se
> sigma.e <- 0.3
> df.1 <- expand.grid(embedding=factor(c("Unembedded","Embedded")),lexicalization=fact
> df <- df.1
> for(i in 1:(n-1))
+   df <- rbind(df,df.1)
> B <- rmvnorm(m,mean=c(0,0),sigma=Sigma.b)
> df$y <- with(df,beta[embedding] + B[cbind(lexicalization,(as.numeric(embedding)))] +
> m <- aov(y ~ embedding + Error(lexicalization/embedding),df)
```



|                         |                                    |                            |
|-------------------------|------------------------------------|----------------------------|
| Embedding<br>134.22     | Lexicalization:Embedding<br>158.34 | Residual<br>Error<br>70.68 |
| Lexicalization<br>49.55 |                                    |                            |

Figure 6.11: Stratification of variance in a simple repeated-measures ANOVA.

### 6.6.6 A case study

This is the outcome of a self-paced reading experiment conducted by Hannah Rohde, in collaboration with me and Andy Kehler.

The question under investigation is whether certain kinds of verbs (*implicit causality* (IC) *verbs*) such as “detest”, which intuitively demand some sort of explanation, can affect readers’ online syntactic attachment preferences.

- (3)
- a. John **detests** the children of the musician who **is** generally arrogant and rude (IC,LOW)
  - b. John **detests** the children of the musician who **are** generally arrogant and rude (IC,HIGH)
  - c. John **babysits** the children of the musician who **is** generally arrogant and rude (NONIC,LOW)
  - d. John **babysits** the children of the musician who **are** generally arrogant and rude (NONIC,HIGH)

We hypothesized that the use of an IC verb should facilitate reading of high-attached RCs, which are generally found in English to be harder to read than low-attached RCs (Cuetos and Mitchell, 1988). The reasoning here is that the IC verbs demand an explanation, and one way of encoding that explanation linguistically is through a relative clause. In these cases, the most plausible type of explanation will involve a clause in which the object of the IC verb plays a role, so an RC modifying the IC verb’s object should become more expected. This stronger expectation may facilitate processing when such an RC is seen (Levy, 2008).

The stimuli for the experiment consist of 20 quadruplets of sentences of the sort above. Such a quadruplet is called an EXPERIMENTAL ITEM in the language of experimental psychol-

ogy. The four different variants of each item are called the CONDITIONS. Since a participant who sees one of the sentences in a given item is liable to be strongly influenced in her reading of another sentence in the item, the convention is only to show each item once to a given participant. To achieve balance, each participant will be shown five items in each condition.

| Participant | Item       |            |            |            |            |     |
|-------------|------------|------------|------------|------------|------------|-----|
|             | 1          | 2          | 3          | 4          | 5          | ... |
| 1           | IC,HIGH    | NONIC,HIGH | IC,LOW     | NONIC,LOW  | IC,HIGH    | ... |
| 2           | NONIC,LOW  | IC,HIGH    | NONIC,HIGH | IC,LOW     | NONIC,LOW  | ... |
| 3           | IC,LOW     | NONIC,LOW  | IC,HIGH    | NONIC,HIGH | IC,LOW     | ... |
| 4           | NONIC,HIGH | IC,LOW     | NONIC,LOW  | IC,HIGH    | NONIC,HIGH | ... |
| 5           | IC,HIGH    | NONIC,HIGH | IC,LOW     | NONIC,LOW  | IC,HIGH    | ... |
| ⋮           | ⋮          | ⋮          | ⋮          | ⋮          | ⋮          | ⋮   |

The experimental data will be analyzed for effects of verb type and attachment level, and more crucially for an *interaction* between these two effects. For this reason, we plan to conduct a two-way ANOVA.

In self-paced reading, the observable effect of difficulty at a given word often shows up a word or two downstream, so in this case we will focus on the first word after the disambiguator—i.e., “generally”. This is called the FIRST SPILLOVER REGION. First we read in the complete dataset, zoom in on the results at this region, and look at the distribution of reading times for each condition. A BOXPLOT (also known as a BOX-AND-WHISKERS diagram) is a good tool for this kind of visualization, and for identifying outliers.<sup>6</sup>

```
> dat1 <- read.table("../data/rohde_levy_kehler_data/results.final.txt",
+   quote = "", sep = "\t", header = T)
> dat1 <- subset(dat1, subj != "subj2" & subj != "subj10" & subj !=
+   "subj50")
> dat1$subj <- factor(dat1$subj)
> spillover.1 <- subset(dat1, expt == 1 & correct == 1 & crit ==
+   "RC_VERB+1")
> spillover.1$verb <- factor(spillover.1$verb)
> spillover.1$attachment <- factor(spillover.1$attachment)
> boxplot(rt ~ verb * attachment, spillover.1)
> boxplot(log(rt) ~ verb * attachment, spillover.1)
```

As you can see in Figure 6.12, there are lots of outliers – though the situation looks better in log-space. Histograms reveal similar results (Figure 6.13):

```
> spillover.1$cond <- factor(spillover.1$cond)
> histogram(~rt | cond, spillover.1, breaks = 30)
```

---

<sup>6</sup>In a boxplot, the upper and lower bounds of the box are the first and third quartile of the data; the length of this box is called the INTER-QUARTILE RANGE, or IQR. The solid-line “whiskers” are placed at the farthest points that lie no more than  $1.5 \times \text{IQR}$  from the edges of the box. Any points that lie beyond these whiskers are considered “outliers” and plotted individually.

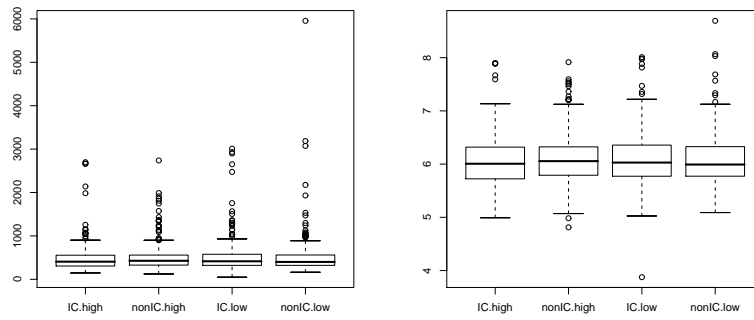


Figure 6.12: Boxplots for raw and log reading times at the first spillover region for the IC-RC experiment

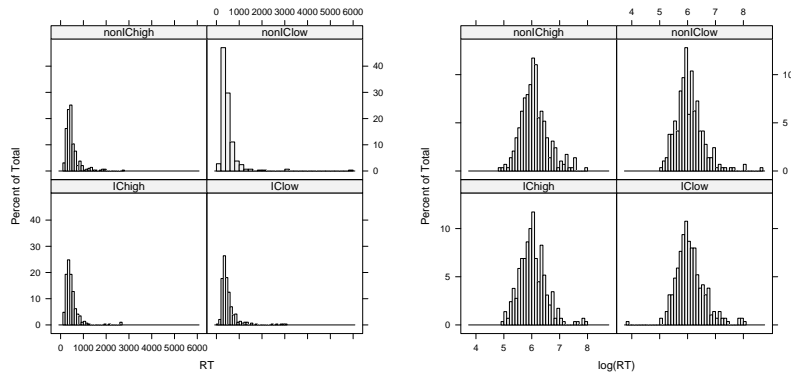


Figure 6.13: Histograms for raw and log reading times at the first spillover region for the IC-RC experiment

It is by no means obvious what to do in this kind of situation where there are so many outliers, and the raw response departs so severely from normality. In the case of self-paced reading, the dominant convention is to perform `OUTLIER REMOVAL`: use some relatively standardized criterion for identifying outliers, and deal with those outliers in some way.

There are several different ways outlier removal is handled in the literature; in this situation we shall apply one of the more common procedures. For each condition, we calculate for each point a `Z-SCORE`, which is a measure of how many sample standard deviations the point lies away from the sample mean. Points with a z-score of magnitude above 4 are simply thrown away:

```
> cbind.list <- function(l) {
+   result <- c()
+   for(i in 1:length(l)) {
+     result <- cbind(result,l[[i]])
+   }
}
```

```

+   result
+ }
> get.z.score <- function(response,conds.list) {
+   means <- tapply(response,conds.list,mean)
+   sds <- tapply(response,conds.list,sd)
+   (response - means[cbind.list(conds.list)]) /
+     sds[cbind.list(conds.list)]
+ }
> z <- with(spillover.1,get.z.score(rt,list(verb,attachment)))
> sum(abs(z) > 4) # 14 points are flagged this way as outliers

[1] 14

> length(z) # 14 out of 933 is a lot of outliers at 4sd!!!!

[1] 933

> # But this is typical for self-paced reading experiments
> spillover.1.to.analyze <- subset(spillover.1,abs(z) <= 4)

```

We take a quick look at what our resulting data look like (some fields of the data frame have been omitted):

```

> head(spillover.1.to.analyze)
  Subj Item Verb Attachment      Crit      RT
9     1   1   IC         high RC_VERB+1 365.27
22    10   1 nonIC        low RC_VERB+1 616.43
35    11   1   IC         low RC_VERB+1 255.56
48    12   1 nonIC        high RC_VERB+1 626.26
61    13   1   IC         high RC_VERB+1 330.45
74    14   1 nonIC        low RC_VERB+1 434.66

```

and now we are ready to conduct our two-way ANOVA.

### The comparisons to make

In this experiment, four factors characterize each stimulus: a particular subject reads a particular item that appears with particular values of the **verb** and **attachment** manipulations. **verb** and **attachment** have two levels each, so if we had  $m$  subjects and  $n$  items we would in principle need at least  $2 \times 2 \times m \times n$  observations to consider a full classic linear model with interactions of all possible types. However, because each subject saw each item only once, we only have  $m \times n$  observations. Therefore it is not possible to construct the full model.

For many years dating back to Clark (1973), the gold standard in this situation has been to construct two separate analyses: one for subjects, and one for items. In the analysis over

| Verb  | Attachment | Subject |       |       |       |       |     |
|-------|------------|---------|-------|-------|-------|-------|-----|
|       |            | 1       | 2     | 3     | 4     | 5     | ... |
| IC    | High       | 280.7   | 396.1 | 561.2 | 339.8 | 546.1 | ... |
|       | Low        | 256.3   | 457.8 | 547.3 | 408.9 | 594.1 | ... |
| nonIC | High       | 340.9   | 507.8 | 786.7 | 369.8 | 453.0 | ... |
|       | Low        | 823.7   | 311.4 | 590.4 | 838.3 | 298.9 | ... |

Table 6.1: Repeated-measures (within-subjects) view of item-aggregated data for subjects ANOVA

subjects, we take as our individual data points the *mean* value of all the observations in each cell of Subject  $\times$  Verb  $\times$  Attachment—that is, we AGGREGATE, or average, across items. Correspondingly, in the analysis over items, we aggregate across subjects. We can use the function `aggregate()` to perform this averaging:

```
sp.1.subj <- with(spillover.1.to.analyze, aggregate(list(rt=rt),
  list(subj=subj, verb=verb, attachment=attachment), mean))
sp.1.item <- with(spillover.1.to.analyze, aggregate(list(rt=rt),
  list(item=item, verb=verb, attachment=attachment), mean))
```

`aggregate()`  
`with()`

The view of the resulting data for the analysis over subjects can be seen in Table 6.1. This setup is called a WITHIN-SUBJECTS or REPEATED-MEASURES design because each subject participates in each condition—or, in another manner of speaking, we take multiple measurements for each subject. Designs in which, for some predictor factor, each subject participates in only one condition are called BETWEEN-SUBJECTS designs.

The way we partition the variance for this type of analysis can be seen in Figure 6.14. Because we have averaged things out so we only have one observation per Subject/Verb/Attachment combination, there will be no variation in the Residual Error box. Each test for an effect of a predictor sets of interest (`verb`, `attachment`, and `verb:attachment`) is performed by comparing the variance explained by the predictor set  $P$  with the variance associated with arbitrary random interactions between the subject and  $P$ . This is equivalent to performing a model comparison between the following two linear models, where  $i$  range over the subjects and  $j$  over the conditions in  $P$ :

$$rt_{ij} = \alpha + B_i \text{Subj}_i + \epsilon_{ij} \quad (\text{null hypothesis}) \quad (6.14)$$

$$rt_{ij} = \alpha + B_i \text{Subj}_i + \beta_j P_j + \epsilon_{ij} \quad (\text{alternative hypothesis}) \quad (6.15)$$

$$(6.16)$$

There is an added wrinkle here, which is that the  $B_i$  are not technically free parameters but rather are themselves assumed to be random and normally distributed. However, this difference does not really affect the picture here. (In a couple of weeks, when we get to mixed-effects models, this difference will become more prominent and we'll learn how to handle it in a cleaner and more unified way.)

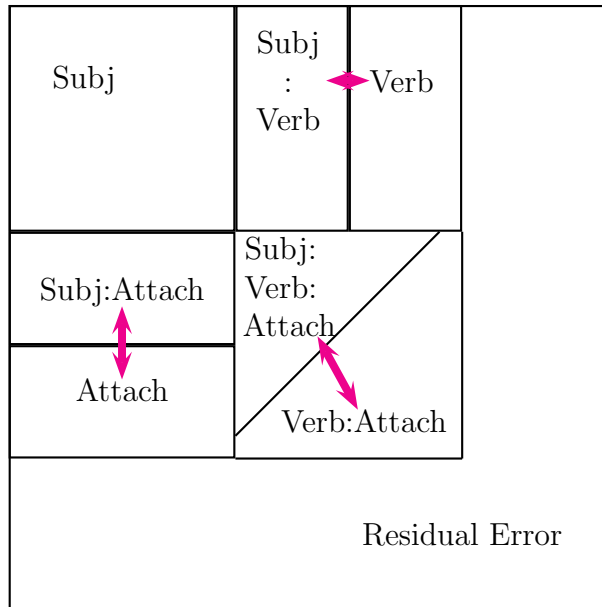


Figure 6.14: The picture for this  $2 \times 2$  ANOVA, where Verb and Attachment are the fixed effects of interest, and subjects are a random factor

Fortunately, `aov()` is smart enough to know to perform all these model comparisons in the appropriate way, by use of the `Error()` specification in your model formula. This is done as follows, for subjects:

```
> summary(aov(rt ~ verb * attachment
+ Error(subj/(verb *attachment)), sp.1.subj))
```

```
Error: subj
      Df Sum Sq Mean Sq F value Pr(>F)
Residuals 54 4063007    75241
```

```
Error: subj:verb
      Df Sum Sq Mean Sq F value Pr(>F)
verb    1  48720   48720  7.0754 0.01027 *
Residuals 54 371834    6886
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Error: subj:attachment
      Df Sum Sq Mean Sq F value Pr(>F)
attachment 1    327    327  0.0406 0.841
Residuals 54 434232    8041
```

```
Error: subj:verb:attachment
              Df Sum Sq Mean Sq F value Pr(>F)
verb:attachment  1  93759   93759  6.8528 0.01146 *
Residuals       54 738819   13682
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

and for items:

```
> summary(aov(rt ~ verb * attachment
+ Error(item/(verb *attachment)), sp.1.item))
```

```
Error: item
              Df Sum Sq Mean Sq F value Pr(>F)
Residuals    19 203631   10717
```

```
Error: item:verb
              Df Sum Sq Mean Sq F value Pr(>F)
verb           1  21181   21181  3.5482 0.075 .
Residuals     19 113419    5969
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Error: item:attachment
              Df Sum Sq Mean Sq F value Pr(>F)
attachment     1    721     721  0.093 0.7637
Residuals     19 147299    7753
```

```
Error: item:verb:attachment
              Df Sum Sq Mean Sq F value Pr(>F)
verb:attachment  1  38211   38211  5.4335 0.03092 *
Residuals       19 133615    7032
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fortunately, the by-subjects and by-items analysis yield largely similar results: they both point towards (a) a significant main effect of verb type; and (b) more interestingly, a significant interaction between verb type and attachment level. To interpret these, we need to look at the means of each condition. It is conventional in psychological experimentation to show the condition means from the aggregated data for the by-subjects analysis:

```
> with(sp.1.subj,tapply(rt,list(verb),mean))
      IC   nonIC
```

```

452.2940 482.0567
> with(sp.1.subj,tapply(rt,list(verb,attachment),mean))
           high      low
IC      430.4316 474.1565
nonIC   501.4824 462.6309

```

The first spillover region was read more quickly in the implicit-causality verb condition than in the non-IC verb condition. The interaction was a **CROSSOVER INTERACTION**: in the high attachment conditions, the first spillover region was read more quickly for IC verbs than for non-IC verbs; but for the low attachment conditions, reading was faster for non-IC verbs than for IC verbs.

We interpreted this result to indicate that IC verbs do indeed facilitate processing of high-attaching RCs, to the extent that this becomes the preferred attachment level.

## 6.7 Other generalized linear models

Recall that we've looked at linear models, which specify a conditional probability density  $P(Y|X)$  of the form

$$Y = \alpha + \beta_1 X_1 + \dots + \beta_n X_n + \epsilon \quad (6.17)$$

Linear models thus assume that the only stochastic part of the data is the normally-distributed noise  $\epsilon$  around the predicted mean. Yet many—probably most—types of data do not meet this assumption at all. These include:

- Continuous data in which noise is not normally distributed;
- Categorical data, where the outcome is one of a number of discrete classes;
- Count data, in which the outcome is restricted to non-negative integers.

By choosing different link and noise functions, you can help ensure that your statistical model is as faithful a reflection of possible of the major patterns in the data you are interested in representing. In the remainder of this chapter, we look at two other major classes of GLM: **LOGIT** and **LOG-LINEAR** models.

### 6.7.1 Logit models

Suppose we want a GLM that models binomially distributed data from  $n$  trials. We will use a slightly different formulation of the binomial distribution from what that of Chapter 2: instead of viewing the response as the number of successful trials  $r$ , we view the response as the *proportion* of successful trials  $\frac{r}{n}$ ; call this  $Y$ . The mean proportion for binomial distribution is simply the success parameter  $\pi$ ; hence,  $\pi$  is also the predicted mean  $\mu$  of our



GLM. This gives us enough information to specify precisely the resulting model (from now on we replace  $\mu$  with  $\pi$  for simplicity):

$$P(Y = y; \pi) = \binom{n}{y} \pi^{ny} (1 - \pi)^{n(1-y)} \quad (\text{or equivalently, replace } \mu \text{ with } \pi) \quad (6.18)$$

which is just the binomial distribution from back in Equation 3.8.

This is the second part of designing a GLM: choosing the distribution over  $Y$ , given the mean  $\mu$  (Equation 6.1). Having done this means that we have placed ourselves in the **BINOMIAL GLM FAMILY**. The other part of specifying our GLM is choosing a relationship between the linear predictor  $\eta$  and the mean  $\mu$ . Unlike the case with the classical linear model, the identity link function is not a possibility, because  $\eta$  can potentially be any real number, whereas the mean proportion  $\mu$  of successes can only vary between 0 and 1. There are many link functions that can be chosen to make this mapping valid, but here we will use the most popular link function, the **LOGIT** transform:<sup>7</sup>

$$\log \frac{\pi}{1 - \pi} = \eta \quad (6.19)$$

or equivalently the **INVERSE LOGIT** transform:

$$\pi = \frac{e^\eta}{1 + e^\eta} \quad (6.20)$$

Figure 6.15 shows the relationship between  $\eta$  and  $\pi$  induced by the logit transform

When we insert the full form of the linear predictor from Equation (6.1) back in, we arrive at the final formula for logit models:

$$\pi = \frac{e^{\alpha + \beta_1 X_1 + \dots + \beta_n X_n}}{1 + e^{\alpha + \beta_1 X_1 + \dots + \beta_n X_n}} \quad (6.21)$$

Fitting a logit model is also called **LOGISTIC REGRESSION**.

## 6.7.2 Fitting a simple logistic regression model

The most common criterion by which a logistic regression model for a dataset is fitted is exactly the way that we chose the parameter estimates for a linear regression model: the method of maximum likelihood. That is, we choose the parameter estimates that give our dataset the highest likelihood.

We will give a simple example using the **dative** dataset. The response variable here is whether the recipient was realized as an NP (i.e., the double-object construction) or as a PP

---

<sup>7</sup>Two other popular link functions for binomial GLMs are the **PROBIT** link and the **COMPLEMENTARY LOG-LOG** link. See Venables and Ripley (2002, Chapter 7) for more details.

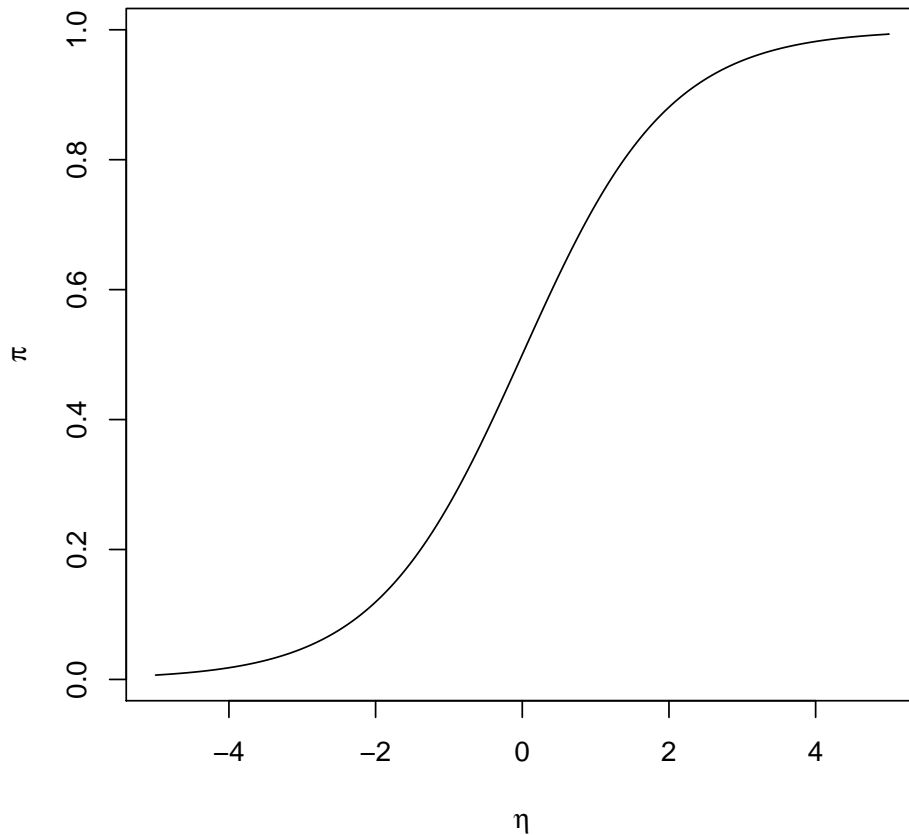


Figure 6.15: The logit transform

(i.e., the prepositional object construction). This corresponds to the `RealizationOfRecipient` variable in the dataset. There are several options in R for fitting basic logistic regression models, including `glm()` in the `stats` package and `lrm()` in the `Design` package. In this case we will use `lrm()`. We will start with a simple study of the effect of recipient pronominality on the dative alternation. Before fitting a model, we examine a contingency table of the outcomes of the two factors:

```
> library(languageR)
> xtabs(~PronomOfRec + RealizationOfRecipient, dative)
```

|               | RealizationOfRecipient |     |
|---------------|------------------------|-----|
| PronomOfRec   | NP                     | PP  |
| nonpronominal | 600                    | 629 |
| pronominal    | 1814                   | 220 |

So sentences with nonpronominal recipients are realized roughly equally often with DO and PO constructions; but sentences with pronominal recipients are recognized nearly 90% of the time with the DO construction. We expect our model to be able to encode these findings.

It is now time to construct the model. To be totally explicit, we will choose ourselves which realization of the recipient counts as a “success” and which counts as a “failure” (although `lrm()` will silently make its own decision if given a factor as a response). In addition, our predictor variable is a factor, so we need to use dummy-variable encoding; we will suffice with the R default of taking the alphabetically first factor level, `nonpronominal`, as the baseline level.

```
> library(Design)
> response <- ifelse(dative$RealizationOfRecipient == "PP", 1,
+ 0)
> lrm(response ~ PronomOfRec, dative)
```

The thing to pay attention to for now is the estimated coefficients for the intercept and the dummy indicator variable for a pronominal recipient. We can use these coefficients to determine the values of the linear predictor  $\eta$  and the predicted mean success rate  $p$  using Equations (6.1) and (6.21):

$$\eta_{--} = 0.0472 + (-2.1569) \times 0 = 0.0472 \quad (\text{non-pronominal recipient}) \quad (6.22)$$

$$\eta_{+} = 0.0472 + (-2.1569) \times 1 = -2.1097 \quad (\text{pronominal recipient}) \quad (6.23)$$

$$p_{\text{nonpron}} = \frac{e^{0.0472}}{1 + e^{0.0472}} = 0.512 \quad (6.24)$$

$$p_{\text{pron}} = \frac{e^{-2.1097}}{1 + e^{-2.1097}} = 0.108 \quad (6.25)$$

When we check these predicted probabilities of PO realization for nonpronominal and pronominal recipients, we see that they are equal to the proportions seen in the corresponding rows of the cross-tabulation we calculated above:  $\frac{629}{629+600} = 0.518$  and  $\frac{220}{220+1814} = 0.108$ . This is exactly the expected behavior, because (a) we have two parameters in our model,  $\alpha$  and  $\beta_1$ , which is enough to encode an arbitrary predicted mean for each of the cells in our current representation of the dataset; and (b) as we have seen before (Section 4.3.1), the maximum-likelihood estimate for a binomial distribution is the relative-frequency estimate—that is, the observed proportion of successes.

### 6.7.3 Multiple logistic regression

Just as we were able to perform multiple linear regression for a linear model with multiple predictors, we can perform multiple logistic regression. Suppose that we want to take into account pronominality of both recipient and theme. First we conduct a complete cross-tabulation and get proportions of PO realization for each combination of pronominality status:

`apply()`

```
> tab <- xtabs(~RealizationOfRecipient + PronomOfRec + PronomOfTheme,
+   dative)
> tab
```

```
, , PronomOfTheme = nonpronominal
```

|                        | PronomOfRec   |            |
|------------------------|---------------|------------|
| RealizationOfRecipient | nonpronominal | pronominal |
| NP                     | 583           | 1676       |
| PP                     | 512           | 71         |

```
, , PronomOfTheme = pronominal
```

|                        | PronomOfRec   |            |
|------------------------|---------------|------------|
| RealizationOfRecipient | nonpronominal | pronominal |
| NP                     | 17            | 138        |
| PP                     | 117           | 149        |

```
> apply(tab, c(2, 3), function(x) x[2]/sum(x))
```

|               | PronomOfTheme |            |
|---------------|---------------|------------|
| PronomOfRec   | nonpronominal | pronominal |
| nonpronominal | 0.4675799     | 0.8731343  |
| pronominal    | 0.0406411     | 0.5191638  |

Pronominality of the theme consistently increases the probability of PO realization; pronominality of the recipient consistently increases the probability of DO realization.

We can construct a logit model with independent effects of theme and recipient pronominality as follows:

```
> library(Design)
> dative.lrm <- lrm(response ~ PronomOfRec + PronomOfTheme, dative)
> dative.lrm
```

And once again, we can calculate the predicted mean success rates for each of the four combinations of predictor variables:

| Recipient | Theme   | $\eta$  | $\hat{p}$ |
|-----------|---------|---------|-----------|
| nonpron   | nonpron | -0.1644 | 0.459     |
| pron      | nonpron | -3.0314 | 0.046     |
| nonpron   | pron    | 2.8125  | 0.943     |
| pron      | pron    | -0.0545 | 0.486     |

In this case, note the predicted proportions of success are not the same as the observed proportions in each of the four cells. This is sensible – we cannot fit four arbitrary means with only three parameters. If we added in an interactive term, we would be able to fit four arbitrary means, and the resulting predicted proportions would be the observed proportions for the four different cells.

| Predictor            | Coefficient | Factor Weight | Multiplicative effect on odds |
|----------------------|-------------|---------------|-------------------------------|
| Intercept            | -0.1644     | 0.4590        | 0.8484                        |
| Pronominal Recipient | -2.8670     | 0.0538        | 0.0569                        |
| Pronominal Theme     | 2.9769      | 0.9515        | 19.627                        |

Table 6.2: Logistic regression coefficients and corresponding factor weights for each predictor variable in the `dative` dataset.

### 6.7.4 Transforming predictor variables

\*\*\*TODO\*\*\*

### 6.7.5 Multiplicativity of the odds

Let us consider the case of a dative construction in which both the recipient and theme are encoded with pronouns. In this situation, both the dummy indicator variables (indicating that the theme and recipient are pronouns) have a value of 1, and thus the linear predictor consists of the sum of three terms. From Equation (6.19), we can take the exponent of both sides and write

$$\frac{p}{1-p} = e^{\alpha+\beta_1+\beta_2} \tag{6.26}$$

$$= e^{\alpha} e^{\beta_1} e^{\beta_2} \tag{6.27}$$

The ratio  $\frac{p}{1-p}$  is the ODDS OF SUCCESS, and in logit models the effect of any predictor variable on the response variable is multiplicative in the odds of success. If a predictor has coefficient  $\beta$  in a logit model, then a unit of that predictor has a multiplicative effect of  $e^{\beta}$  on the odds of success.

Unlike the raw coefficient  $\beta$ , the quantity  $e^{\beta}$  is not linearly symmetric—it falls in the range  $(0, \infty)$ . However, we can also perform the full REVERSE LOGIT TRANSFORM of Equation (6.20), mapping  $\beta$  to  $\frac{e^{\beta}}{1+e^{\beta}}$  which ranges between zero and 1, and is linearly symmetric around 0.5. The use of logistic regression with the reverse logit transform has been used in quantitative sociolinguistics since Cedergren and Sankoff (1974) (see also Sankoff and Labov, 1979), and is still in widespread use in that field. In quantitative sociolinguistics, the use of logistic regression is often called VARBRUL (variable rule) analysis, and the parameter estimates are reported in the reverse logit transform, typically being called FACTOR WEIGHTS.

Tables 6.2 and 6.3 show the relationship between the components of the linear predictor, the components of the multiplicative odds, and the resulting predictions for each possible combination of our predictor variables.

| Recip. | Theme | Linear Predictor              | Multiplicative odds                     | P(PO) |
|--------|-------|-------------------------------|---|-------|
| -pron  | -pron | -0.16                         | 0.8484                                  | 0.46  |
| +pron  | -pron | $-0.16 - 2.87 = -3.03$        | $0.85 \times 0.06 = 0.049$              | 0.046 |
| -pron  | +pron | $-0.16 + 2.98 = 2.81$         | $0.85 \times 19.6 = 16.7$               | 0.94  |
| +pron  | +pron | $-0.16 - 2.87 + 2.98 = -0.05$ | $0.85 \times 0.06 \times 19.63 = 0.947$ | 0.49  |

Table 6.3: Linear predictor, multiplicative odds, and predicted values for each combination of recipient and theme pronominality in the `dative` dataset. In each case, the linear predictor is the log of the multiplicative odds.

## 6.8 Confidence intervals and model comparison in logit models

We'll close our introduction to logistic regression with discussion of confidence intervals and model comparison.

### 6.8.1 Frequentist Confidence intervals for logit models

When there are a relatively large number of observations in comparison with the number of parameters estimated, the standardized deviation of the MLE for a logit model parameter  $\theta$  is approximately normally distributed:

$$\frac{\hat{\theta} - \theta}{\text{StdErr}(\hat{\theta})} \sim \mathcal{N}(0, 1) \quad (\text{approximately}) \quad (6.28)$$

This is called the WALD STATISTIC<sup>8</sup>. This is very similar to the case where we used the  $t$  statistic for confidence intervals in classic linear regression (Section 6.4; remember that once the  $t$  distribution has a fair number of degrees of freedom, it basically looks like a standard normal distribution). If we look again at the output of the logit model we fitted in the previous section, we see the standard error, which allows us to construct confidence intervals on our model parameters.

|                          | Coef    | S.E.    | Wald Z | P      |
|--------------------------|---------|---------|--------|--------|
| Intercept                | -0.1644 | 0.05999 | -2.74  | 0.0061 |
| PronomOfRec=pronominal   | -2.8670 | 0.12278 | -23.35 | 0.0000 |
| PronomOfTheme=pronominal | 2.9769  | 0.15069 | 19.75  | 0.0000 |

Following the exact same logic as in Section 6.4, we find that the 95% confidence interval for each parameter  $\beta_i$  is bounded below by  $\hat{\beta}_i - 1.96SE(\hat{\beta}_i)$ , and bounded above by  $\hat{\beta}_i + 1.96SE(\hat{\beta}_i)$ . This gives us the following bounds:

---

<sup>8</sup>It is also sometimes called the Wald Z statistic, because of the convention that standard normal variables are often denoted with a Z, and the Wald statistic is distributed approximately as a standard normal.

```
a -0.1673002  0.2762782
b1 -3.1076138 -2.6263766
b2  2.6815861  3.2722645
```

The Wald statistic can also be used for a frequentist test on the null hypothesis that an individual model parameter is 0. This is the source of the  $p$ -values given for the model parameters above.

### 6.8.2 Bayesian confidence intervals for logit models

In order to construct a Bayesian confidence interval for a logit model, we need to choose prior distributions on the weights  $\alpha$  and  $\{\beta_i\}$  that go into the linear predictor (Equation (6.1)), and then use sampling-based techniques (Section 4.5). As a simple example, let us take the multiple logistic regression of Section 6.7.3. The model has three parameters; we will express agnosticism about likely parameter values by using a diffuse prior. Specifically, we choose a normally-distributed prior with large variance for each parameter:

$$\begin{aligned}\alpha &\sim \mathcal{N}(0, 10000) \\ \beta_1 &\sim \mathcal{N}(0, 10000) \\ \beta_2 &\sim \mathcal{N}(0, 10000)\end{aligned}$$

With sampling we can recover 95% HPD confidence intervals (Section 5.1) for the parameters:

```
a -0.1951817  0.2278135
b1 -3.1047508 -2.6440788
b2  2.7211833  3.2962744
```

There is large agreement between the frequentist and Bayesian confidence intervals in this case. A different choice of prior would change the HPD confidence intervals, but we have a lot of data relative to the complexity of the model we're trying to estimate, so the data dominates the prior in our case.

### 6.8.3 Model comparison

Just as in the analysis of variance, we are often interested in conducting tests of the hypothesis that introducing *several* model parameters simultaneously leads to a better overall model. In this case, we cannot simply use a single Wald statistic for hypothesis testing. Instead, the most common approach is to use the LIKELIHOOD-RATIO TEST, first introduced in Section 5.4.4. To review, the quantity

$$G^2 = 2 [\log \text{Lik}_{M_1}(\mathbf{y}) - \log \text{Lik}_{M_0}(\mathbf{y})] \tag{6.29}$$

is approximately distributed as a  $\chi_k^2$  random variable, where  $k$  is the difference in the number of free parameters between  $M_1$  and  $M_0$ .

As an example of using the likelihood ratio test, we will hypothesize a model in which pronominality of theme and recipient both still have additive effects but that these effects may vary depending on the modality (spoken versus written) of the dataset. We fit this model and our modality-independent model using `glm()`, and use `anova()` to calculate the `glm()` likelihood ratio:

```
> m.0 <- glm(response ~ PronomOfRec + PronomOfTheme, dative, family = "binomial")
> m.A <- glm(response ~ PronomOfRec * Modality + PronomOfTheme *
+   Modality, dative, family = "binomial")
> anova(m.0, m.A)
```

We can look up the  $p$ -value of this deviance result in the  $\chi_3^2$  distribution:

```
> 1 - pchisq(9.07, 3)
```

```
[1] 0.02837453
```

Thus there is some evidence that we should reject a model that doesn't include modality-specific effects of recipient and theme pronominality.

#### 6.8.4 Dealing with symmetric outcomes

In the study of language, there are some types of categorical outcomes that are symmetrical in a way that can make it difficult to see how to properly assign values to explanatory variables. Consider, for example, the study of word order in the coordination of like categories. Suppose we are interested in the joint effect of word frequency and word length on ordering preferences in word pairs conjoined by *and* (called, appropriately enough, BINOMIALS), and our observation is the phrase *evasive and shifty*. The word *evasive* is longer (has more syllables) than *shifty*, but it is less frequent as well. How do we characterize these independent variables, and do we call the outcome a “success” or a “failure”?

Fortunately, we can address this problem by noticing that the central issue is really not whether *evasive and shifty* is a success or failure; the central issue is, rather, the pattern of how the explanatory variables are aligned with observed orderings. We now cover an example of how to deal with this problem taken from Benor and Levy (2006), a corpus study of English binomials. We will restrict ourselves to word pairs occurring exactly once in Benor and Levy's dataset, and look at the effects of perceptual markedness, weight (in terms of number of syllables), and word frequency. The covariates in the model are thus comparative properties—for example, whether one of the words denotes a property that is more perceptually salient, or which of the words is more frequent (also *chanted*). We can code each property  $P_i$  as a quantitative variable  $X_i$  by arbitrarily choosing an alignment direction for the property, and giving the binomial a positive value for the  $X_i$  if  $P_i$  is aligned with the binomial, a negative value of equal magnitude if  $P_i$  is aligned against the binomial,



and zero if  $P_i$  is inactive. The logit response variable now serves as a dummy variable—it is always a “success”. For perceptual markedness, word length, and word frequency we choose the following alignments:

- Perceptual markedness is positive if the first word in the binomial is more perceptually salient than the last word;
- Word length (in number of syllables) is positive if the last word has more syllables than the first word;
- Word frequency is positive if the first word is more frequent than the last word.

These aligned properties can be thought of as SOFT (or GRADIENT) CONSTRAINTS in the sense of Optimality Theory and similar frameworks, with statistical model fitting as a principled means of investigating whether the constraints tend not to be violated, and how strong such a tendency may be. A few such observations in the dataset thus coded are:

| Word.Pair            | Percept | Freq | Syl | Response |
|----------------------|---------|------|-----|----------|
| chanted and chortled | 1       | 1    | 0   | 1        |
| real and vibrant     | 0       | -1   | -1  | 1        |
| evasive and shifty   | 0       | 1    | -1  | 1        |

Note that *chanted and chortled* has a perceptual markedness value of 1, since chortling is a quieter action; *vibrant and real* has a response of 0 since it is observed in the opposite ordering; and the Syl covariate value for *evasive and shifty* is  $-1$  because *evasive* has more syllables than *shifty*.

It would be nonsensical to use an intercept when fitting a model to this dataset: setting the intercept arbitrarily high, and the other model parameters to zero, would be the best fit. If, however, we remove the intercept from the model, the model expresses the tendency of each covariate to align with the binomial ordering:

```
> dat <- read.table("../data/binomials_data/single_count_binomials.txt",
+   header = T, fill = T, as.is = T)
> summary(glm(Response ~ Percept + Syl + Freq - 1, dat, family = "binomial"))$coef
```

|         | Estimate  | Std. Error | z value  | Pr(> z )    |
|---------|-----------|------------|----------|-------------|
| Percept | 1.1771339 | 0.5158658  | 2.281861 | 0.022497563 |
| Syl     | 0.4926385 | 0.1554392  | 3.169332 | 0.001527896 |
| Freq    | 0.3660976 | 0.1238079  | 2.956981 | 0.003106676 |

All three constraints have positive coefficients, indicating significant alignment with binomial ordering: the constraints do indeed tend not to be violated. It’s worth noting that even though perceptual markedness is estimated to be the strongest of the three constraints (largest coefficient), its standard error is also the largest: this is because the constraint is active (non-zero) least often in the dataset.

## 6.9 Log-linear and multinomial logit models

A class of GLM very closely related to logit models is LOG-LINEAR MODELS. Log-linear models choose the log as the link function:

$$l(\mu) = \log \mu = \eta \qquad \mu = e^\eta \qquad (6.30)$$

and the Poisson distribution, which ranges over non-negative integers, as the noise function:

$$P(Y = y; \mu) = e^{-\mu} \frac{\mu^y}{y!} \quad (y = 0, 1, \dots) \qquad (6.31)$$

When used to model count data, this type of GLM is often called a POISSON MODEL or POISSON REGRESSION.

In linguistics, the log-linear model is most often used to model probability distributions over multi-class outcomes. Suppose that there are  $M$  classes of possible outcomes, each with its own linear predictor  $\eta_i$  and random variable  $Y_i$ . If we conditionalize on the total count of all classes being 1, then the only available count outcomes for each class are 0 and 1, with probabilities:

$$P(Y_i = 1; \mu_i = e^{\eta_i}) = e^{\mu_i} e_i^\eta \qquad P(Y_i = 0; \mu_i = e^{\eta_i}) = e^{-\mu_i} \qquad (6.32)$$

and the joint probability of the single observation falling into class  $i$  is

$$\begin{aligned} P(Y_i = 1, \{Y_{j \neq i}\} = 0) &= \frac{e^{\mu_i} e_i^\eta \prod_{j \neq i} e^{\mu_j}}{\sum_{i'} e^{\mu_{i'}} e_{i'}^\eta \prod_{j \neq i'} e^{\mu_j}} \\ &= \frac{e_i^\eta \prod_j e^{\mu_j}}{\sum_{i'} e_{i'}^\eta \prod_j e^{\mu_j}} \\ &= \frac{e_i^\eta \prod_j e^{\mu_j}}{\prod_j e^{\mu_j} \sum_{i'} e_{i'}^\eta} \\ P(Y_i = 1, \{Y_{j \neq i}\} = 0) &= \frac{e_i^\eta}{\sum_{i'} e_{i'}^\eta} \end{aligned} \qquad (6.33)$$

When we are thinking of a log-linear model as defining the probability distribution over which class each observation falls into, it is often useful to define the class-specific success probabilities  $\pi_i \stackrel{\text{def}}{=} P(Y_i = 1, \{Y_{j \neq i}\} = 0)$ . This allows us to think of a log-linear model as using a MULTINOMIAL noise distribution (Section 3.4.1).

## Expressive subsumption of (multinomial) logit models by log-linear models\*

Basic logit models are used to specify probability distributions over outcomes in two classes (the “failure” class 0 and the “success” class 1). Log-linear models can be used to specify probability distributions over outcomes in any number of classes. For a two-class log-linear model, the success probability for class 1 is (Equation (6.21)):

$$\pi_1 = \frac{e^{\alpha_1 + \beta_{1,1}X_1 + \dots + \beta_{1,n}X_n}}{e^{\alpha_0 + \beta_{0,1}X_1 + \dots + \beta_{0,n}X_n} + e^{\alpha_1 + \beta_{1,1}X_1 + \dots + \beta_{1,n}X_n}} \quad (6.34)$$

$$(6.35)$$

If we divide both the numerator and denominator by  $e^{\alpha_0 + \beta_{0,1}X_1 + \dots + \beta_{0,n}X_n}$ , we get

$$\pi_1 = \frac{e^{(\alpha_1 - \alpha_0) + (\beta_{1,1} - \beta_{0,1})X_1 + \dots + (\beta_{1,n} - \beta_{0,n})X_n}}{1 + e^{(\alpha_1 - \alpha_0) + (\beta_{1,1} - \beta_{0,1})X_1 + \dots + (\beta_{1,n} - \beta_{0,n})X_n}} \quad (6.36)$$

This is significant because the model now has exactly the same form as the logit model (Equation (6.21)), except that we have parameters of the form  $(\alpha_1 - \alpha_0)$  and  $(\beta_{1,i} - \beta_{0,i})$  rather than  $\alpha$  and  $\beta_i$  respectively. This means that log-linear models EXPRESSIVELY SUBSUME logit models: any logit model can also be expressed by some log-linear model. Because of this, when maximum-likelihood estimation is used to fit a logit model and a log-linear model with the same set of variables, the resulting models will determine the same probability distribution over class proportions. There are only three differences:

1. The log-linear model can also predict the total number of observations.
2. The logit model has fewer parameters.
3. When techniques other than MLE (e.g., Bayesian inference marginalizing over model parameters) are used, the models will generally yield different predictive distributions.

## 6.10 Log-linear models of phonotactics

We introduce the framework of LOG-LINEAR or MAXIMUM-ENTROPY models by turning to the linguistic problem of phonotactics. A speaker’s PHONOTACTIC KNOWLEDGE is their knowledge of what logically possible sound sequences constitute legitimate potential lexical items in her language. In the probabilistic setting, phonotactic knowledge can be expressed as a probability distribution over possible sound sequences. A good probabilistic model of phonotactics assigns low probability to sequences that are not possible lexical items in the language, and higher probability to sequences that are possible lexical items. A categorical characterization of sound sequences as being either impossible or possible in the language could be identified with respective assignment of zero or non-zero probability in the model. The classic example of such a distinction is that whereas native English speakers judge the

non-word *blick* [blik] to be a possible word of English, they judge the non-word *bnick* [bnik] not to be a possible word of English. [citations here] However, probabilistic phonotactic models have the further advantage of being able to make *gradient* distinctions between forms that that are “more” or “less” appropriate as possible lexical items.

Construct a probabilistic phonotactic model entails putting a probability distribution over the possible sound sequences of the language. There are many approaches that could be taken to this problem; here we examine two different approaches in the context of modeling one of the best-studied problems in phonotactics: constraints on of English *word onsets*—the consonant sequences with which words begin. For simplicity, we restrict discussion here to onsets consisting of exactly two segments drawn from a subset of the inventory of English consonants, namely [f], [v], [s], [z], [sh], [p], [b], [t], [d], [l], and [r]. Table 6.4 presents a list of the two-segment word onsets which can be constructed from these segments and which are found in the Carnegie Mellon Pronouncing Dictionary of English (Weide, 1998). Of the 121 logically possible onsets, only 30 are found. They are highly disparate in frequency, and most of the rarest (including everything on the right-hand side of the table except for [sf] as in *sphere*) are found only in loan words. In the study of phonotactics; there is some question as to exactly what counts as an “attested” sequence in the lexicon; for present purposes, I will refer to the twelve most frequent onsets plus [sf] as *unambiguously attested*.

We begin the problem of estimating a probability distribution over English two-segment onsets using simple tools from Chapters 2 and 4: multinomial models and relative frequency estimation. Let us explicitly represent the sequence structure of an English onset  $x_1x_2$  as a  $\#_Lx_1x_2\#_R$ , where  $\#_L$  represents the left edge of the onset and  $\#_R$  represents the right edge of the onset. Every two-segment onset can be thought of as a linearly ordered joint event comprised of the left edge, the first segment, the second segment, and the right edge. We can use the chain rule to represent this joint event as a product of conditional probabilities:

$$P(\#_Lx_1x_2\#_R) = P(\#_L)P(x_1|\#_L)P(x_2|\#_Lx_1)P(\#_R|\#_Lx_1x_2) \quad (6.37)$$

The left edge is obligatory, so that  $P(\#_L) = 1$ ; and since we are restricting our attention to two-segment onsets, the right edge is also obligatory when it occurs, so that  $P(\#_R|\#_Lx_1x_2) = 1$ . We can thus rewrite Equation 6.37 as

$$P(\#_Lx_1x_2\#_R) = P(x_1|\#_L)P(x_2|\#_Lx_1) \quad (6.38)$$

We consider three possible methods for estimating this probability distribution from our data:

1. Treat each complete onset  $\#_Lx_1x_2\#_R$  as a single outcome in a multinomial model, with 121 possible outcomes; the problem then becomes estimating the parameters of this single multinomial from our data. As described in Chapter XXX, the maximum likelihood estimate for multinomials is also the relative frequency estimate, so the

probability assigned to an onset in this model is directly proportional to the onset's frequency of occurrence.

With this model it is also useful to note that for any segment  $x$ , if the event  $y$  immediately preceding it is not the left edge  $\#_L$ , then  $y$  itself is preceded by  $\#_L$ . This means that  $P(x_2|\#_Lx_1) = P(x_2|x_1)$ . This allows us to rewrite Equation ??:

$$P(\#_Lx_1x_2\#_R) = P(x_1|\#_L)P(x_2|x_1) \quad (6.39)$$

Hence this model can also be thought of as a *bigram* model in which the probability of an event is, given the immediately preceding event, conditionally independent on everything earlier in the sequence. Note here that if we have  $N$  possible segments, we must fit  $N + 1$  multinomial distributions.

2. We can introduce the strong independence assumption that the probability of a segment is entirely independent of its context:  $P(x_i|x_{1\dots i-1}) = P(x_i)$ . This is a *unigram* model, giving

$$P(\#_Lx_1x_2\#_R) = P(x_1)P(x_2) \quad (6.40)$$

Here we need to fit only one multinomial distribution.

3. We can introduce the somewhat weaker independence assumption that the probability of a segment depends on whether it is the first or second segment in the onset, but not on what other segments occur in the onset.

$$P(\#_Lx_1x_2\#_R) = P(x_1|\#_L)P(x_2|\#_L\_) \quad (6.41)$$

where  $\_$  indicates the presence of *any* segment. We might call this a *positional unigram* model to emphasize the position-dependence. This model requires that we fit two multinomial distributions.

Columns 3–5 of Table 6.4 show estimated probabilities for attested onsets in these three models. Major differences among the models are immediately apparent. Among unambiguously attested onsets, [st] is much more probable in the bigram model than in either unigram model; [tr] and [sf] are much more probable in the unigram model than in the other two models; and [sp] is much less probable in the positional unigram model (see also Exercise 6.11).

A substantive claim about the nature of phonotactic knowledge put forth by researchers including Hayes and Wilson (2007) as well as XXX is that probabilistic models which do a good job accounting for the distribution of segment sequences in the lexicon should also

| Segment | Freq | $P_{\text{unigram}}$ | $P_{\text{unipos}}$ | $P_{\text{bigram}}$ | Segment | Freq | $P_{\text{unigram}}$ | $P_{\text{unipos}}$ | $P_{\text{bigram}}$ |
|---------|------|----------------------|---------------------|---------------------|---------|------|----------------------|---------------------|---------------------|
| st      | 1784 | 0.0817               | 0.0498              | 0.1755              | vl      | 15   | 0.0011               | 0.0006              | 0.0015              |
| br      | 1500 | 0.1122               | 0.112               | 0.1476              | vr      | 14   | 0.003                | 0.0016              | 0.0014              |
| pr      | 1494 | 0.1405               | 0.1044              | 0.147               | sf      | 12   | 0.0395               | 0.0003              | 0.0012              |
| tr      | 1093 | 0.1555               | 0.0599              | 0.1075              | sr      | 10   | 0.1553               | 0.154               | 0.001               |
| fr      | 819  | 0.0751               | 0.0745              | 0.0806              | zl      | 9    | 0.0003               | 0.0003              | 0.0009              |
| sp      | 674  | 0.0738               | 0.0188              | 0.0663              | zb      | 4    | 0.0003               | 0.0                 | 0.0004              |
| bl      | 593  | 0.0428               | 0.0427              | 0.0583              | sht     | 4    | 0.0067               | 0.0041              | 0.0004              |
| fl      | 572  | 0.0286               | 0.0284              | 0.0563              | dv      | 3    | 0.0002               | 0.0001              | 0.0003              |
| pl      | 458  | 0.0535               | 0.0398              | 0.0451              | zv      | 2    | 0.0                  | 0.0                 | 0.0002              |
| dr      | 441  | 0.0239               | 0.0239              | 0.0434              | tv      | 2    | 0.0016               | 0.0003              | 0.0002              |
| sl      | 379  | 0.0592               | 0.0587              | 0.0373              | dz      | 2    | 0.0001               | 0.0                 | 0.0002              |
| shr     | 155  | 0.0128               | 0.0128              | 0.0152              | tl      | 1    | 0.0593               | 0.0228              | 0.0001              |
| shl     | 79   | 0.0049               | 0.0049              | 0.0078              | shv     | 1    | 0.0001               | 0.0001              | 0.0001              |
| ts      | 23   | 0.0817               | 0.0003              | 0.0023              | sb      | 1    | 0.059                | 0.0001              | 0.0001              |
| sv      | 19   | 0.0016               | 0.0008              | 0.0019              | fs      | 1    | 0.0395               | 0.0003              | 0.0001              |

Table 6.4: The attested two-segment onsets of English, based on the segments [f], [v], [s], [z], [sh], [p], [b], [t], [d], [l], and [r], sorted by onset frequency. Probabilities are relative frequency estimates, rounded to 4 decimal places.

be able to accurately predict native-speaker judgments of the acceptability of “nonce” words (sequences that are not actually words) such as *blick* and *bnick* as potential words of the language. Challenges for this approach become apparent when one examines existing datasets of native-speaker nonce-word judgments. For example, Scholes (1966) conducted a study of English onsets in nonce-word positions and uncovered regularities which seem challenging for the multinomial models we considered above. Among other results, Scholes found the following differences between onsets in the frequency with which nonce words containing them were judged acceptable:

$$(4) \quad [\text{br}] > [\text{vr}] > [\text{sr}], [\text{ml}] > [\text{sf}] > [\text{zl}], [\text{fs}] > [\text{zv}]$$

The fact that the unattested onset [ml] leads to greater acceptability than the unambiguously attested onset [sf] clearly indicates that English phonotactic knowledge involves *some* sorts of generalization beyond the raw contents of the lexicon; hence the bigram model of Table 6.4 is unacceptable. At the same time, however, [br] is clearly preferred to [sr], indicating that both unigram models are too simplistic. One might consider a mixture model which interpolates between bigram and unigram models. The difficulty with this approach, however, is that no simple mixture is obvious that would achieve the preferences necessary. The preference of [sr] over [sf] would seem to indicate that unigrams should receive considerable weighting; but the preference of [vr] over [sr] would be undermined by heavy unigram weighting.

To motivate our next development, let us consider specifically the mystery of the relative acceptability of [vr] and [sr] among onsets that are not unambiguously attested. A key piece of information we have not yet considered is the phonological substructure of the segments in

|         |                                       |                                      |                                     |                      |                  |
|---------|---------------------------------------|--------------------------------------|-------------------------------------|----------------------|------------------|
| Place   | Labial<br>[p],[b]                     | Labiodental<br>[f],[v]               | Alveolar<br>[s],[z],[t],[d],[r],[l] | Alveopalatal<br>[sh] | Velar<br>[k],[g] |
| Manner  | Stop<br>[p],[b],[t],[d],[k],[g]       | Fricative<br>[s],[z],[f],[v],[sh]    | Liquid<br>[r],[l]                   |                      |                  |
| Voicing | Voiced<br>[b],[d],[g],[v],[z],[r],[l] | Unvoiced<br>[p],[t],[k],[f],[s],[sh] |                                     |                      |                  |

Table 6.5: Simple phonological decomposition of the consonants used in Table 6.4

question. There are many ways of representing phonological substructure, but one straightforward approach for consonants is a representation that decomposes each segment into three PHONOLOGICAL FEATURES: its PLACE of articulation, MANNER of articulation, and VOICING [refs]. The value of each of these features for each consonant used in our current example can be found in Table 6.5. The set of segments picked out by some conjunction of phonological features or their exclusion is often called a NATURAL CLASS. For example, among the consonants currently under consideration, the phonological feature [+labial] picks out the natural class {[p],[b]}; the feature [-stop] picks out {[s],[z],[f],[v],[sh],[r],[l]}; the phonological feature conjunction [+labiodental,-voiced] picks out the natural class {[f]}; and so forth.

### 6.10.1 Log-linear models

With multinomial models, it is not obvious how one might take advantage of the featural decomposition of segments in constructing a probability distribution over the discrete set of possible phoneme sequences. We now turn to a modeling framework that allows such decompositions to be taken into account in modeling such discrete random variables: the framework of LOG-LINEAR models. In this framework, which is intimately related to the logistic-regression models covered previously (see Section XXX), the goal is once again modeling conditional probability distributions of the form  $P(Y|X)$ , where  $Y$  ranges over a countable set of response *classes*  $\{y_i\}$ . Unlike the cases covered previously in this chapter, however, the log-linear framework is relatively agnostic to the representation of  $X$  itself. What is crucial, however, is the presence of a finite set of FEATURE FUNCTIONS  $f_j(X, Y)$ , each of which maps every possible paired instance of  $X$  and  $Y$  to a real number. Taken in aggregate, the feature functions map each possible response class  $y_i$  to a FEATURE VECTOR  $\langle f_1(x, y_i), f_2(x, y_i), \dots, f_n(x, y_i) \rangle$ . Finally, each feature function  $f_j$  has a corresponding parameter  $\lambda_j$ . Given a collection of feature functions, corresponding parameter values, and a value  $x$  for the conditioning random variable  $X$ , the conditional probability of each class  $y_i$  is defined to be:

$$P(Y = y_i | X = x) = \frac{1}{Z} \exp \left[ \sum_j \lambda_j f_j(x, y_i) \right] \quad (6.42)$$

where  $Z$  is a normalizing term ensuring that the probability distribution is proper.

In order to translate our phonotactic learning problem into the log-linear framework, we must identify what serves as the conditioning variable  $X$ , the response  $Y$ , and what the feature functions  $f_i$  are. Since we are putting a probability distribution over logically possible English onsets, the response must be which onset found in a (possible) lexical item. The feature functions should correspond to the phonological features identified earlier.<sup>9</sup> Finally, since we are only trying to fit a single probability distribution over possible English onsets that is not dependent on any other information, whatever we take the conditioning variable  $X$  to be, our feature functions will not depend on it; so we can simplify our problem somewhat so that it involves fitting the distribution  $P(Y)$  using feature functions  $f_j(Y)$  with parameters  $\lambda_j$  (also called FEATURE WEIGHTS), with the functional form of the probability distribution as follows:

$$P(Y = y_i) = \frac{1}{Z} \exp \left[ \sum_j \lambda_j f_j(y_i) \right] \quad (6.43)$$

### Simple log-linear models of English onsets

What remains is for us to choose the feature functions for our phonotactic model. This choice of feature functions determines what generalizations can be directly encoded in our model. As a first, highly oversimplified model, we will construct exactly one feature function for each natural class specifiable by a single phonological feature of *manner* or *voicing*. This feature function will return the number of segments in that natural class contained in the onset. That is,

$$f_j(y_i) = \begin{cases} 2 & \text{if both segments in onset } i \text{ belong to the } j\text{-th natural class;} \\ 1 & \text{if only one segment in onset } i \text{ belongs to the } j\text{-th natural class;} \\ 0 & \text{if neither segment in onset } i \text{ belongs to the } j\text{-th natural class.} \end{cases} \quad (6.44)$$

There are four manner/voicing phonological features for our segment inventory; each can be negated, giving eight natural classes.<sup>10</sup> Each onset is thus mapped to an eight-dimensional feature vector. In the onset [sr], for example we would have the following counts:

| Natural class | Matching segments in [sr] | Natural class | Matching segments in [sr] |
|---------------|---------------------------|---------------|---------------------------|
| [+stop]       | 0                         | [-stop]       | 2                         |
| [+fric]       | 1                         | [-fric]       | 1                         |
| [+liquid]     | 1                         | [-liquid]     | 1                         |
| [+voiced]     | 1                         | [-voiced]     | 1                         |

---

<sup>9</sup>Note that the term *feature* is being used in two different here: on the one hand, as part of a decomposition of individual phonological segments, on the other hand as a function that will apply to entire onsets and which is associated with a parameter in the log-linear model. Although phonological features could be used directly as features in the log-linear model, the space of possible log-linear model features is much richer than this.

<sup>10</sup>We omit the phonological feature of unvoicedness because, since voicing here is a binary distinction, [+unvoiced] would be equivalent to [-voiced].



so that the feature vector for [sr] in this model would be  $\langle 0, 2, 1, 1, 1, 1, 1, 1 \rangle$ .

What remains is for us to fit the parameter values  $\lambda_1, \dots, \lambda_8$  corresponding to each of these features. For a simple model like this, in which there are relatively few parameters (eight) for many outcome classes (121) and many observations (10,164), maximum likelihood estimation is generally quite reliable. We find the following maximum-likelihood estimates for our eight feature weights:

|           |         |           |         |
|-----------|---------|-----------|---------|
| [+stop]   | -0.0712 | [-stop]   | 0.0928  |
| [+fric]   | -0.5472 | [-fric]   | 0.5012  |
| [+liquid] | 0.5837  | [-liquid] | -0.6679 |
| [+voiced] | -0.4713 | [-voiced] | 0.7404  |

Similar to the case in logistic regression, positive feature weights indicates preference for onsets with large values for the feature in question, and negative feature weights indicate dispreference for such onsets. The model has learned that stops are slightly dispreferred to non-stops; fricatives and liquids are strongly preferred to non-fricatives and non-liquids; and unvoiced consonants are strongly preferred to voiced consonants. A sharper view, however, of the generalizations made by the model can be seen in Table 6.6, which shows the probabilities placed by this model on attested onsets. Although there are some things that seem to be correct about this model's generalizations—for example, none of the unambiguously attested onsets are given probability below 0.004—the model makes far too few distinctions, leading to problems such as the assignment of high probability to [tʌ], and the assignment of identical probabilities to [ts] and [st]. This failure should have been expected, however, given that our feature functions failed to encode any positional information, or to distinguish at all between certain segments, such as [t] and [p].

We address some of these concerns by moving on to a more complex model, which allows the following generalizations as feature functions:

- Preferences for particular segments to occur in position 1;
- Preferences for particular segments to occur in position 2;
- Preferences for particular bigrams of natural classes specifiable by a single phonological feature of either manner or voicing.

The first two types of features give the log-linear model the same generalizational power as the positional unigram model we covered earlier. The third type of feature, however, goes beyond the positional unigram model, and allows the model to make use of abstract phonological features in generalizing over possible lexical items. Formally speaking, we have one feature function for each segment in position 1, one feature function for each segment in position 2, and one feature function for each possible pairing of single-feature natural classes. This gives us twenty-two possible single-segment feature functions and  $8 \times 8 = 64$  possible bigram feature functions, for a total of 86. We will let each serve as an INDICATOR FUNCTION mapping those onsets it correctly describes to the value 1, and all other onsets to the value 0:

| Onset | Freq | $P_{M_1}$ | $P_{M_2}$ | $P_{M_3A}$ | $P_{M_3B}$ | Onset | Freq | $P_{M_1}$ | $P_{M_2}$ | $P_{M_3A}$ | $P_{M_3B}$ |
|-------|------|-----------|-----------|------------|------------|-------|------|-----------|-----------|------------|------------|
| st    | 1784 | 0.0097    | 0.1122    | 0.1753     | 0.1587     | vl    | 15   | 0.0035    | 0.0007    | 0.0013     | 0.0018     |
| br    | 1500 | 0.0086    | 0.1487    | 0.1473     | 0.1415     | vr    | 14   | 0.0035    | 0.0018    | 0.0014     | 0.0034     |
| pr    | 1494 | 0.0287    | 0.1379    | 0.1468     | 0.1442     | sf    | 12   | 0.004     | 0.0003    | 0.0009     | 0.001      |
| tr    | 1093 | 0.0287    | 0.0791    | 0.1075     | 0.1033     | sr    | 10   | 0.0119    | 0.0915    | 0.0014     | 0.0155     |
| fr    | 819  | 0.0119    | 0.0443    | 0.0802     | 0.0726     | zl    | 9    | 0.0035    | 0.0004    | 0.0009     | 0.0017     |
| sp    | 674  | 0.0097    | 0.0423    | 0.066      | 0.056      | zb    | 4    | 0.0009    | 0.0001    | 0.0001     | 0.0001     |
| bl    | 593  | 0.0086    | 0.0567    | 0.0582     | 0.0541     | sht   | 4    | 0.0097    | 0.0093    | 0.0003     | 0.0039     |
| fl    | 572  | 0.0119    | 0.0169    | 0.0561     | 0.0454     | dv    | 3    | 0.0009    | 0.0001    | 0.0001     | 0.0001     |
| pl    | 458  | 0.0287    | 0.0526    | 0.045      | 0.046      | zv    | 2    | 0.0004    | 0         | 0.0001     | 0          |
| dr    | 441  | 0.0086    | 0.0317    | 0.0432     | 0.0391     | tv    | 2    | 0.0029    | 0.0001    | 0.0001     | 0.0002     |
| sl    | 379  | 0.0119    | 0.0349    | 0.0374     | 0.0359     | dz    | 2    | 0.0009    | 0         | 0          | 0.0001     |
| shr   | 155  | 0.0119    | 0.0076    | 0.0153     | 0.0143     | tl    | 1    | 0.0287    | 0.0301    | 0.0006     | 0.0106     |
| shl   | 79   | 0.0119    | 0.0029    | 0.0077     | 0.0067     | shv   | 1    | 0.0012    | 0.0001    | 0          | 0          |
| ts    | 23   | 0.0097    | 0.0005    | 0.0017     | 0.0002     | sb    | 1    | 0.0029    | 0.0002    | 0.0002     | 0.0016     |
| sv    | 19   | 0.0012    | 0.0011    | 0.0017     | 0.0002     | fs    | 1    | 0.004     | 0.0003    | 0.0001     | 0.0005     |

Table 6.6: Probabilities estimated from four log-linear models for attested English onsets consisting of pairs from the segment inventory [f], [v], [s], [z], [sh], [p], [b], [t], [d], [l].

$$f_j(y_i) = \begin{cases} 1 & \text{if the } j\text{-th feature describes } y_i; \\ 0 & \text{otherwise.} \end{cases} \quad (6.45)$$

As a concrete example of how these feature functions would be applied, let us again consider the onset [sr]. It satisfies the following descriptions:

- [s] is in position 1 (we represent this feature as  $s.$ , with  $.$  indicating that anything can appear in position 2)
- [r] is in position 2 (we represent this feature as  $.r$ )
- [-liquid][+voice]
- All pairwise combinations of [-liquid],[-stop],[+fric],[-voice] in position 1 with [+liquid],[-stop],[-fric],[+voice] in position (16 combinations in total)

Thus the feature vector for [sr] would have eighteen entries of 1 and 68 entries of 0. Using the method of maximum likelihood to estimate values for the 86 parameters of the model, we find that the features with strongest absolute preferences and dispreferences are as follows:

|                    |                         |
|--------------------|-------------------------|
| [-voice][-voice]   | 5.62962436676025390625  |
| .v                 | 3.64033722877502441406  |
| [-liquid][-stop]   | 2.91994524002075195312  |
| [-voice][-stop]    | 2.36018443107604980469  |
| s.                 | 1.81566941738128662109  |
| [-liquid][+liquid] | 1.72637474536895751953  |
| .t                 | 1.68454444408416748047  |
| [-stop][-liquid]   | 1.56158518791198730469  |
| ...                |                         |
| .b                 | -1.03666257858276367188 |
| z.                 | -1.07121777534484863281 |
| [-liquid][-liquid] | -1.20901763439178466797 |
| [-stop][+fric]     | -1.24043428897857666016 |
| .f                 | -1.30032265186309814453 |
| [-stop][-stop]     | -1.85031402111053466797 |
| .d                 | -1.97170710563659667969 |
| .sh                | -3.09503102302551269531 |

Brief inspection indicates that most of the model's strongest preferences involve generalization on natural class cooccurrences: preference for onsets to involve pairs of unvoiced segments, dispreference for pairs matching in manner of articulation, and so forth. In addition, some strong segment-specific positional dispreferences are also found, such as the preference for initial [s] and dispreference for initial [sh]. Caution is required, however, in interpreting individual feature weights too simplistically—for example, it is clear from the lexicon of English that [sh] is dispreferred even more strongly in the second position than in the first position, yet second-position [sh] feature does not appear in the list of most strongly dispreferred features. The reason for this is that several other features—including the four with the largest negative weights—strongly penalize second-position [sh] already. As with linear and logistic regression models, the proper interpretation of a feature weight is what effect a change in the associated feature value would have, *if all other feature values were kept constant*.

The other way of inspecting the generalizations made by the model is by looking at the predictive distribution on the response variable itself, as seen in Table 6.6. This model has a number of clear advantages over our simplest model: it is relatively successful at giving unambiguously attested onsets higher probability than attested onsets, but at the same time gives [sr] higher probability than many other onsets, including some that are unambiguously attested. However, it also has some weaknesses: for example, the probability for [sf] has dropped below many onsets that are not unambiguously attested, such as [vl].

### Overparameterization and regularization

At some level, we might want to allow our model to have specific sensitivity to the frequency of *every* possible onset, so that each instance of a given onset  $x_1x_2$  contributes directly

and idiosyncratically to the probability of other words with that onset; but at the same time, we clearly want our model to generalize to onsets that do not occur in the English lexicon as well. Within the maximum-likelihood log-linear framework we have developed thus far, these two requirements are in conflict with one another, for the following reason. In order to allow the model sensitivity to the frequency of specific onsets, we would want to introduce one feature function for each possible onsets, giving us 121 feature functions and thus 121 parameters to estimate. However, this parameterization allows the encoding of *any* probability distribution over the 121 possible response classes. As we saw in Chapter 4, the maximum-likelihood estimate for a multinomial distribution is just the relative-frequency estimate. Hence a maximum-likelihood log-linear model with onset-specific feature functions would simply memorize the relative frequencies of the onsets. Since adding more natural class-based features to the model can only *increase* its expressivity, no ML-estimated model with these 121 features will generalize beyond relative frequencies.

It *is* possible, however, to learn both onset-specific knowledge and natural-class-level generalizations simultaneously within the log-linear framework, however, by moving away from maximum-likelihood point estimation and instead adopting a Bayesian framework. Recall that in the Bayesian approach, the posterior probability of the model parameters  $\lambda$  is proportional to the likelihood of the data under  $\lambda$  times the prior probability of  $\lambda$ , which in our case works out to:

$$P(\lambda|Y) = P(Y|\lambda)P(\lambda) \tag{6.46}$$

$$= \left[ \prod_i \frac{1}{Z} \exp \left[ \sum_j \lambda_j f_j(y_i) \right] \right] P(\lambda) \tag{6.47}$$

with  $Z$  a normalizing factor dependent on  $\lambda$ . The next step is to choose a prior distribution over the model parameters,  $P(\lambda)$ . In principle, any prior could be used; in practice, a popular choice is a multivariate *Gaussian* prior distribution (Section 3.5) with center  $\mu$  and covariance matrix  $\Sigma$ , so that the prior probability of an  $N$ -dimensional parameter vector  $\lambda$  is

$$P(\lambda) = \frac{1}{\sqrt{(2\pi|\Sigma|)^N}} \exp \left[ \frac{(\lambda - \mu)^T \Sigma^{-1} (\lambda - \mu)}{2} \right] \tag{6.48}$$

This choice of prior is popular for three reasons: (i) it has an intuitive interpretation as encoding a bias toward the parameter vector  $\mu$  that is weak in the vicinity of  $\mu$  but grows rapidly stronger with increasing distance from  $\mu$ ; (ii) for log-linear models, the posterior distribution over  $\lambda$  remains convex with a Gaussian prior; and (iii) Gaussian priors have been found to work well in allowing fine-grained learning while avoiding overfitting with log-linear models. The simplest choice of prior is one in which with mean  $\mu = 0$  and a diagonal covari-

ance matrix whose nonzero entries are all the same value:  $\Sigma = \begin{pmatrix} \sigma^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma^2 \end{pmatrix}$ . Multivariate

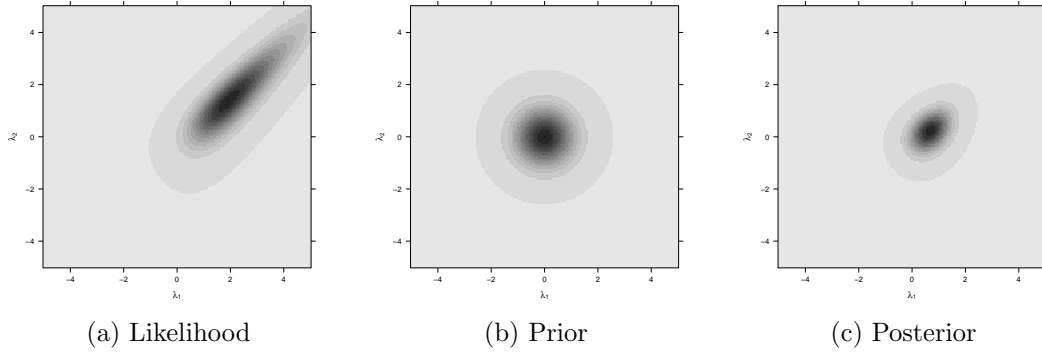


Figure 6.16: A multivariate Gaussian prior ( $\mu = 0, \sigma = 1$ ) for a simple log-linear model with three possible response classes and two indicator features functions:  $f_1$  associated with class 1 and  $f_2$  associated with class 2. First panel is model likelihood for class 1–3 frequencies of 7, 4, and 1 respectively; second panel is the prior distribution; third panel is the posterior distribution.

Gaussian distributions like this are often called SPHERICAL, because surfaces of equal probability are (hyper-)spheres. With a spherical Gaussian prior, the posterior distribution can be written as follows:

$$P(\lambda|Y) \propto P(Y|\lambda)P(\lambda) \tag{6.49}$$

$$= \left[ \prod_i \frac{1}{Z} \exp \left[ \sum_j \lambda_j f_j(y_i) \right] \right] \exp \left[ \sum_j \frac{-\lambda_j^2}{2\sigma^2} \right] \tag{6.50}$$

If we shift to log space we get

$$\log P(\lambda|Y) \propto \overbrace{\sum_i \left[ \log \frac{1}{Z} + \sum_j \lambda_j f_j(y_i) \right]}^{\text{Log-likelihood}} - \overbrace{\left[ \sum_j \frac{\lambda_j^2}{2\sigma^2} \right]}^{\text{Negative log-prior probability}} \tag{6.51}$$

Note that the log-posterior probability falls off quadratically with the sum of the feature weights. For this reason, a Gaussian prior is sometimes called a QUADRATIC prior.

The effect of a Gaussian prior of this form can be seen in Figure 6.16: the prior penalizes deviations from its mode of 0 (a symmetric model in which all outcomes are equally likely), so that the posterior mode falls in between the MLE and the prior mode.

Let us now turn back to our study of English onsets, ready to apply our Bayesian log-linear model. We are now in a position to deploy a richer set of feature functions: on top of the positional single-segment and paired natural-class features we included in the previous model, we add paired-segment and positional single-natural-class features. This gives us an inventory of 223 total feature functions; the feature-vector representation for the onset

[sr], for example, would now have the paired-segment feature **sr**, as well as the positional single-natural-class features [-stop]., [+fric]., [-liquid]., [-voiced]., [-stop], [-fric], [+liquid], and [+voiced].

As is always the case with Bayesian inference, we have a number of choices as to handle the problems of parameter estimation and prediction. Unlike the case with multinomial models, however, there are no readily available analytic techniques for dealing with Bayesian log-linear models, and sampling techniques can be quite computationally intensive. A popular approach is to use maximum a-posteriori (MAP) estimation to find the set of feature weights with (near-)maximum posterior probability, and to approximate Bayesian prediction by using these MAP parameter estimates. In our problem, using a symmetric Gaussian prior centered around  $\mathbf{0}$  with standard deviation  $\sigma = 1$ , the features with largest and smallest weights in the MAP estimate are as follows:

|                  |                         |
|------------------|-------------------------|
| st               | 3.24827671051025390625  |
| sp               | 2.78049993515014648438  |
| fl               | 2.51510095596313476562  |
| ts               | 2.09755825996398925781  |
| s.               | 1.87449419498443603516  |
| [-voice][-voice] | 1.80559206008911132812  |
| fr               | 1.80392193794250488281  |
| .v               | 1.72066390514373779297  |
| ...              |                         |
| [-stop][+fric]   | -0.67749273777008056641 |
| [+voice][-voice] | -0.70867472887039184570 |
| .d               | -1.00191879272460937500 |
| shp              | -1.00633215904235839844 |
| ss               | -1.12540340423583984375 |
| fp               | -1.57261526584625244141 |
| .sh              | -1.64139485359191894531 |
| zr               | -1.65136361122131347656 |
| ft               | -1.85411751270294189453 |
| dl               | -2.24138593673706054688 |
| tl               | -3.16438293457031250000 |
| sr               | -4.12058639526367187500 |

Comparison with the previous model indicates important overall similarities, but it is clear that the new features are also being used by the model, perhaps most notably in encoding idiosyncratic preferences for [st] and dispreferences for [sr] and [sh]. The predictive distribution of this model,  $M_{3A}$ , can be found in Table 6.6. As expected, there is more probability mass on unambiguously attested onsets in this model than in either previous model, since this model is able to directly encode idiosyncratic preferences for specific onsets. Additionally, much of the apparent weakness of  $M_2$  has been partly remedied—for example, the probability of [sr] has dropped below all the other unambiguously attested sequences except for [sf] while the lower probability of [vr] has stayed about the same.

## Strength of the prior and generalization in log-linear models

What is the effect of increasing the strength of the prior distribution, as encoded by decreasing the standard deviation  $\sigma$  of the spherical multivariate Gaussian? There are two key effects we'll cover here consideration. The first effect is an overall tendency for the posterior to look more like the prior, a straightforward and intuitive consequence of the fact that in Bayesian inference, prior and likelihood stand on equal ground in determining posterior beliefs. There is a second, more subtle effect that merits attention, however, and which becomes clear from careful inspection of Equation 6.51. Consider the contribution of an individual feature weight  $\lambda_j$  to the posterior probability of the complete parameter vector  $\lambda$ . The choice of  $\lambda_j$  contributes directly to the log-likelihood once for *every* observation for which the corresponding feature is implicated, but contributes to the prior log-probability only once regardless of how many observations in which the corresponding feature is implicated. This fact leads has an important consequence: a stronger prior penalizes feature weights more heavily the *sparser* the corresponding feature—that is, the less often that feature is unambiguously implicated in the data.

We can illustrate this consequence by re-fitting our previous log-linear phonotactic model using a much stronger prior: a spherical Gaussian distribution with standard deviation  $\sigma = 0.01$ . The resulting probability distribution over attested onsets is shown in Table 6.6 as model  $M_{3B}$ . Compared with  $M_{3A}$  (which had  $\sigma = 1$ ), there is an overall shift of probability mass away from unambiguously attested onsets; this is the first effect described above. However, the remaining onsets do *not* all undergo similar increases in probability: the onsets [sr] and [tl], for example, undergo very large increases, whereas onsets such as [vl] and [zb] stay about the same. The reason for this is as follows. The more general features—natural-class and segment unigrams and natural-class bigrams—favor [sr] and [tl]: in our data, [s] and [t] are common as the first segment of two-segment onsets, [r] and [l] are common as the second segment of two-segment onsets, and [-voiced][+liquid] is a common natural-class bigram. The burden of fitting the low empirical frequency of [sr] and [tl] falls on the most specific features—segment bigrams—but large weights for specific features are disfavored by the strong prior, so that the resulting predictive probabilities of these onsets rises. In contrast, [vl] and [zb] are not favored by the more general features, so that their predictive probability does not rise appreciably with this moderate increase in prior strength.

### A word of caution

Finally, a word of caution is necessary in the practical use of MAP estimation techniques with overparameterized log-linear models: even using Bayesian techniques so that the MAP estimate is well-defined, the posterior distribution can be *very* flat in the vicinity of its optimum, which can make it difficult to be sure how close the obtained solution may be to the true optimum. In these cases, one would do well to impose stringent convergence criteria on whatever optimization algorithm is used to search for the MAP estimate.

## Log-linear distributions are maximum-entropy distributions

\*mention the term maxent, and point out that log-linear models satisfy the maxent property\*

### 6.10.2 Translating between logit models and log-linear models

Although we have demonstrated in Section 6.9 that log-linear models expressively subsume logit models, translating between the two can require some care. We go through a brief example here.

\*\*\*SAY MORE\*\*\*

Gabe's needs doing example.

```
> dat <- read.table("../data/needs_doing_data/needs.txt", header = T)
> dat$Response <- ifelse(dat$Response == "ing", 1, 0)
> dat$Anim1 <- factor(ifelse(dat$Anim == "abst", "abst", "conc"))
> model.logit <- glm(Response ~ Anim1 + sqrt(Dep.Length), data = dat,
+   family = binomial)
> dat.for.loglin <- with(dat, as.data.frame(as.table(tapply(Response,
+   list(Anim1 = Anim1, Dep.Length = Dep.Length, Response = Response),
+   length))))
> names(dat.for.loglin)[4] <- "x"
> dat.for.loglin$DL <- dat.for.loglin$Dep.Length
> dat.for.loglin$Dep.Length <- as.numeric(as.character(dat.for.loglin$DL))
> dat.for.loglin$Response <- as.numeric(as.character(dat.for.loglin$Response))
> dat.for.loglin$x <- sapply(dat.for.loglin$x, function(x) ifelse(is.na(x),
+   0, x))
> model.loglin <- glm(x ~ Anim1 * DL + Response + Response:(Anim1 +
+   sqrt(Dep.Length)), data = dat.for.loglin, family = "poisson")
> summary(model.loglin)$coef[c(32, 62, 63), ]
```

|                           | Estimate   | Std. Error | z value   | Pr(> z )     |
|---------------------------|------------|------------|-----------|--------------|
| Response                  | -0.2950173 | 0.1107085  | -2.664812 | 7.703144e-03 |
| Anim1conc:Response        | 1.3333414  | 0.1431564  | 9.313880  | 1.232457e-20 |
| Response:sqrt(Dep.Length) | -0.6048434 | 0.0636931  | -9.496215 | 2.176582e-21 |

```
> summary(model.logit)$coef
```

|                  | Estimate   | Std. Error | z value   | Pr(> z )     |
|------------------|------------|------------|-----------|--------------|
| (Intercept)      | -0.2950173 | 0.11070848 | -2.664812 | 7.703141e-03 |
| Anim1conc        | 1.3333414  | 0.14315636 | 9.313881  | 1.232446e-20 |
| sqrt(Dep.Length) | -0.6048434 | 0.06369308 | -9.496218 | 2.176519e-21 |

What we see here is that the “effect of the response variable category” in the log-linear model corresponds to the intercept in the logit model; and the interactions of response with



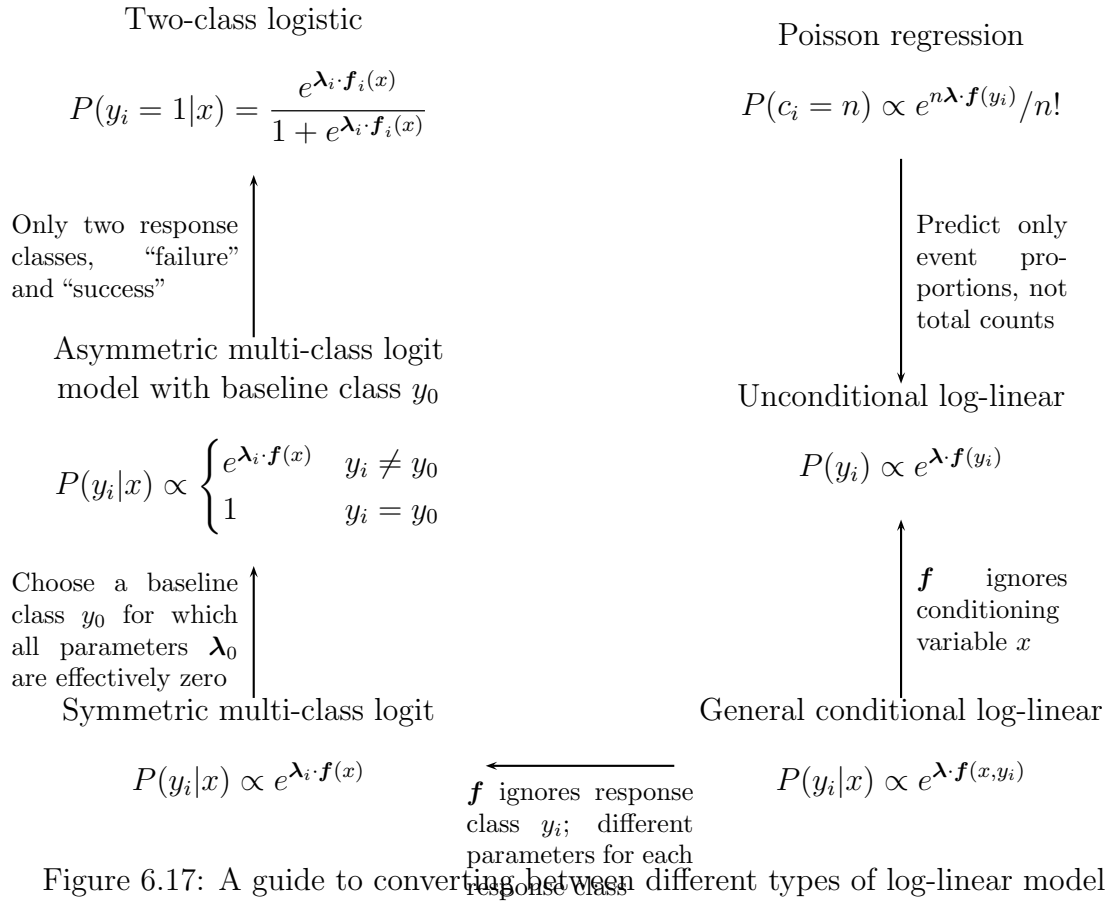


Figure 6.17: A guide to converting between different types of log-linear models

animacy and dependency length in the log-linear model correspond to the animacy and dependency length effects in the logit model. Of course, the logit model is far more efficient to fit; it involved only three parameters, whereas the log-linear model required sixty-three.

\*\*\*WHAT ABOUT MODELS WHERE WE HAVE NO BASELINE CLASS BUT ALSO DON'T NEED ALL THOSE EXTRA PARAMETERS TO MODEL THE COUNTS?\*\*\*

...

## 6.11 Guide to different kinds of log-linear models

Because we have covered several types of log-linear models in this chapter, it is useful to take a moment to carefully consider the relationship among them. A diagram making these relationships explicit is given in Figure 6.17. This section briefly describes these relationships. For brevity, we have used dot-product notation instead of summation notation: model parameters and feature-function outcomes are both denoted with vectors  $\lambda$  and  $f(x, y_i)$ , so that the weighted sums  $\sum_j \lambda_j f_j(x, y_i)$  we have seen previously can be succinctly expressed as dot products  $\lambda \cdot f(x, y_i)$ .

In the bottom-right corner of Figure 6.17 is the general conditional log-linear model we

covered in Section XXX. In this model, there is a collection of feature functions  $f_j$  each of which maps an input  $x$  paired with a response class  $y_i$  to a real number. Each feature function  $f_j$  has an associated parameter weight  $\lambda_j$ . In the general conditional log-linear model, no further constraints are placed on the nature of these feature functions.

It is a common modeling decision, however, to assume that there should effectively be a single set of feature functions shared identically by all possible response classes. As an example, consider the problem of relativizer choice for non-subject extracted relative clauses with animate head nouns, such as in *the actress — you mentioned*. In modeling this problem with conditional distributions  $P(\text{Relativizer}|\text{Context})$ , one might consider three possible response classes: *that*, *who(m)*, and relativizer omission. To examine the effect of frequency of the head noun (here, *actress*), we might want to come up with a single numerical encoding (say, log of Brown-corpus frequency) which is associated with a different feature function for each response class. Thus we would have three feature functions  $f_{1,2,3}$ , each of which is defined as follows:

$$f_j(x, y_i) = \begin{cases} \text{Log head-noun frequency} & j = i \\ 0 & \text{otherwise} \end{cases}$$

An equivalent approach, however, would be to say that there is only one feature function  $f_1$  which always returns log head-noun frequency, but has a different parameter  $\lambda_{1i}$  for each response class. Taking this approach moves us to the lower-left corner of Figure 6.17: symmetric multi-class logistic regression. This category of model is more restrictive than the general conditional log-linear model: the latter can express probability distributions unavailable to the former, by using feature functions that are active for more than one response class, or by using feature functions active for one response class which have no matching feature function for another response class.

Some readers may have noticed that the symmetric multi-class logit model has more parameters than it needs. Let us identify one of the  $N$  response classes  $y_0$  as the BASELINE CLASS. Then for an input  $x$ , we can the probability of any outcome  $y_i$  is as follows:

$$P(y_i|x) = \frac{e^{\lambda_i \mathbf{f}(x)}}{e^{\lambda_0 \mathbf{f}(x)} + e^{\lambda_1 \mathbf{f}(x)} + \dots + e^{\lambda_N \mathbf{f}(x)}} \quad (6.52)$$

Let us now divide both the top and bottom of this fraction by  $e^{\lambda_0 \mathbf{f}(x)}$ :

$$P(y_i|x) = \frac{e^{\lambda_i \mathbf{f}(x)} \frac{1}{e^{\lambda_0 \mathbf{f}(x)}}}{[e^{\lambda_0 \mathbf{f}(x)} + e^{\lambda_1 \mathbf{f}(x)} + \dots + e^{\lambda_N \mathbf{f}(x)}] \frac{1}{e^{\lambda_0 \mathbf{f}(x)}}}} \quad (6.53)$$

$$= \frac{e^{[\lambda_i - \lambda_0] \mathbf{f}(x)}}{e^{[\lambda_0 - \lambda_0] \mathbf{f}(x)} + e^{[\lambda_1 - \lambda_0] \mathbf{f}(x)} + \dots + e^{[\lambda_N - \lambda_0] \mathbf{f}(x)}}} \quad (6.54)$$

But  $\lambda_i - \lambda_0 = \mathbf{0}$ , so  $e^{[\lambda_0 - \lambda_0] \mathbf{f}(x)} = 1$ . If we now define  $\lambda'_i \equiv \lambda_i - \lambda_0$ , we have:

$$P(y_i|x) = \frac{e^{\lambda'_i \mathbf{f}(x)}}{1 + e^{\lambda'_1 \mathbf{f}(x)} + \dots + e^{\lambda'_N \mathbf{f}(x)}} \quad (6.55)$$

This is a new expression of the same model, but with fewer parameters. Expressing things in this way leads us to the middle-left model in Figure 6.17. This is an *asymmetric* multiclass logit model in that we had to distinguish one class as the “baseline”, but it is just as expressive as the symmetric multiclass logit model: any probability distribution that can be represented with one can be represented with the other. Therefore, any predictive inferences made using maximum-likelihood estimation techniques will be the same for the two approaches. Other techniques—such as Bayesian MAP parameter estimation or Bayesian prediction while “integrating out”—may lead to different results, however, due to the sensitivity of the prior to the structure of model parameterization.

Cases of this model where there are only two possible outcome classes are traditional *two-class* logit models (top left corner of Figure 6.17), which we covered in detail in Section XXX. This is the type of log-linear model that the majority of readers are likely to have encountered first.

Returning to the general conditional log-linear case in the bottom-right corner of Figure 6.17, another option is to omit any sensitivity of feature functions to the input  $x$ . This is equivalent to throwing out the conditioning-variable part of the model altogether, and is sensible in cases such as our modeling of phonotactic knowledge (Section XXX), where simply wanted to infer a single probability distribution over English two-segment onsets. This decision takes us to the middle-right cell in Figure 6.17, UNCONDITIONAL log-linear models.

Finally, the unconditional log-linear model that we have here is closely related to another type of generalized linear model: POISSON REGRESSION. The key difference between unconditional log-linear models as we have described them here and Poisson regression is as follows: whereas our models have placed multinomial distributions over a set of possible response classes, the goal of Poisson regression is to put a probability distribution over *counts of observed events* in each possible response class. The two models are intimately related: if we take a fitted Poisson-regression model and use it to compute the joint probability distribution over counts in response class subject to the constraint that the total count of all response classes is 1, we get the same probability distribution that would be obtained using an unconditional log-linear model with the same parameters (Exercise 6.17). Although Poisson regression is popular in statistical modeling in general, we have not covered it here; it does turn up in some work on language modeling the frequencies of event counts in large corpora (e.g., Baayen, 2001).

## 6.12 Feed-forward neural networks

XXX

## 6.13 Further reading

There are many places to go for reading more about generalized linear models and logistic regression in particular. The classic comprehensive reference on generalized linear models

is McCullagh and Nelder (1989). For GLMs on categorical data, Agresti (2002) and the more introductory Agresti (2007) are highly recommended. For more information specific to the use of GLMs and logistic regression in R, Venables and Ripley (2002, Section 7), Harrell (2001, Chapters 10–12), and Maindonald and Braun (2007, Section 8.2) are all good places to look.

## 6.14 Notes and references

There are many good implementations of log-linear/maximum-entropy models publicly available; one that is simple to use from the command line, flexible, and fast is MegaM (Daumé and Marcu, 2006).

- Mention  $L_1$  prior in addition to  $L_2$  prior.

## 6.15 Exercises

### Exercise 6.1: Linear regression

1. The dataset `english` contains lexical-decision reaction times for 2284 English words along with a number of properties of each word. Use linear regression to assess the relationship between NEIGHBORHOOD DENSITY (defined as the number of words of English differing from the target word by only a single-letter edit) and reaction time. Is higher neighborhood density associated with faster or slower decision times? Introduce written word (log-)frequency as a control variable. Does the direction of the neighborhood-density effect change? Is it a reliable effect (that is, what is its level of statistical significance)?
2. The dataset `nonwordsLexdec` presents average reaction times for 39 *non-word* letter sequences of English in a primed lexical decision experiment by Bicknell et al. (2010). The prime preceding the non-word always *was* a word, so trials were of the form *dish-kess*, *otter-peme*, and so forth. The dataset also contains neighborhood densities for each of the non-words, and word log-frequencies for the primes. Use linear regression to assess the relationship between neighborhood density and lexical-decision reaction time, controlling for prime log-frequency. Is the relationship between neighborhood density and reaction time the same as for the `english` dataset? Is the relationship reliable? Why do you see the results you see?

### Exercise 6.2: Linear regression

The `durationsGe` dataset has as dependent variable the length of the Dutch prefix *ge-* in seconds. Use linear regression to investigate which of the following predictors have significant effects on prefix length:

- Word frequency
- Speaker sex
- Speech rate

Make sure to account for the possibility of interactions between the predictors. In addition, for word frequency and speech rate, use data visualization and `loess()` to get an intuition for whether to transform the predictors before putting them in the regression. (**Hint:** to get rid of rows in a data frame with NA's in them, the function `is.na()` is useful.)

### Exercise 6.3: Analysis of variance

We talked about the idea of using a log-transformation on response variables such as reaction times to make them look more normal and hence be more faithful to the assumptions of linear models. Now suppose you are conducting a two-way ANOVA and are interested in the possibility of an interaction between the two factors. Your data are reaction times and look more normal when log-transformed. What are the potential consequences of log-transforming your response variable for investigating whether there is an interaction between your two predictors of interest? **Hint:** try constructing a set of four condition means for a two-by-two that reflect an additive pattern, and then look at the pattern when you take the log of each cell.

### Exercise 6.4: Linear regression

Compare the residualization and multiple linear regression approaches. Imagine an underlying model of reading time of words in sentences in which the negative logs of raw word frequency ( $F_{log}$ ) and contextual predictability ( $P_{log}$ ) both play a role in determining the average reading time ( $RT$ , measured in milliseconds) of a given word. Take as the model of average reading time

$$RT = 300 - 50F_{log} - 10P_{log} + \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 40)$$

and suppose that  $F_{log}$  and  $P_{log}$  are generated from a multivariate normal distribution centered at  $(-4, -4)$  with variance-covariance matrix  $\begin{pmatrix} 0.7 & 0.5 \\ 0.5 & 1.2 \end{pmatrix}$ . In this case where predictability and frequency are positively correlated, is your intuition that residualization or multiple linear regression will have greater statistical power in detecting the effect of predictability? (That is, on average which approach will yield a higher proportion of successful detections of a significance effect of predictability?) Test your intuitions by comparing residualization versus multiple linear regression approaches for detecting the effect of  $P_{log}$ . Generate 1000 sample datasets, each of size 200. Which approach has more statistical power in detecting the effect of predictability? **Hint:** You can automatically extract a  $p$ -value from the  $t$ -statistic for a regression model parameter by looking at the fourth component of the `summary()` of an `lm` object (the result of `summary()` is a list), which is an array. For example:



$\beta = 0$  against an alternative-hypothesis model  $M_A$  with unconstrained  $\beta$  by comparing the variance explained by  $M_A$  over  $M_0$  with the residual variance unexplained by  $M_A$ , will in general lead to anti-conservative inference. That is: assume  $\beta = 0$ ; choose values of  $\alpha$ ,  $\Sigma$ , and  $\sigma^2$ , the number of speakers  $m > 1$  and utterances per speaker  $n > 1$ ; randomly generate  $N$  datasets using this model; analyze each dataset using a non-repeated-measures procedure; and report the proportion of models in which the null hypothesis would be rejected by the criterion  $p < 0.05$ .

2. Now demonstrate that the stratification-of-error procedure introduced in Section 6.6.5 avoids anti-conservative inference, through repeated generation of simulated data as in the first part of this problem.

### Exercise 6.8: Analysis of variance.

Perform by-subjects and by-items repeated-measures ANOVA analyses of the second spillover region (RC\_VERB+2) the ? self-paced reading dataset. Try the results both with and without applying outlier removal; a typical outlier-removal criterion would be to discard observations more than either three or four standard deviations above the mean, with “mean” and “standard deviation” determined using only observations from the specific region being analyzed. How, if at all, does outlier removal change the results? Why do you think this is the case?

### Exercise 6.9: The $t$ -test versus Bayesian model comparison

Consider data that is generated from two normal distributions, with means  $\mu_1 = 1$  and  $\mu_2 = 2.5$ , and with common noise  $\sigma_\epsilon = 5$ . Let’s look at the power of the frequentist  $t$ -test versus a Bayesian model comparison in choosing between hypotheses  $H_0$  in which the two distributions have the same mean, versus  $H_1$  in which the two distributions have different means. Assume that our observations  $Y$  consist of 250 points from each distribution. For the Bayesian model comparison, use the specifications

$$\begin{aligned}\mu_1 &\sim \mathcal{N}(0, \sigma_\mu) \\ \mu_2 - \mu_1 &\sim \mathcal{N}(0, \sigma_\mu) \\ \sigma_\epsilon &\sim \mathcal{U}(1, 100)\end{aligned}$$

and the prior distribution  $P(H_0) = P(H_1) = 1/2$ . Using JAGS or similar sampling-based Bayesian inference software, plot the proportion of trials in which the posterior probability of  $H_0$  is less than 0.05:  $P(H_0|Y) < 0.05$ , as a function of  $\sigma_\mu$ . Explain the pattern you see in intuitive terms.

### Exercise 6.10: Logistic regression

In Lecture 13, in analyzing the `datave` dataset we found in a logit model with linear predictor

$$\text{RealizationOfRecipient} \sim \text{PronomOfRec} + \text{PronomOfTheme}$$

that pronominality of recipient and theme had similar-sized but opposite effects (in logit space) on the probability of use of the prepositional-object construction. We tentatively interpreted this result as consistent with the idea that there is a general “pronouns like to be shifted left” constraint that operates with equal strength on recipients and themes.

1. The model above (call it  $M_1$ ) has three free parameters. Define a new predictor variable that (a) is a function of the two variables `PronomOfRec` and `PronomOfTheme`; and (b) allows us to simplify the model above into a new model with only two free parameters.
2. Fit the model (call it  $M_2$ ) to the `dative` dataset. How do the resulting parameter estimates compare to those of  $M_1$ ?
3. Your new model  $M_2$  should be nested inside  $M_1$  (that is, it should be a special case of  $M_1$ ). Explain this nesting—specifically, explain what special conditions imposed on  $M_1$  result in equivalence to  $M_2$ . This nesting makes it possible to conduct a likelihood-ratio test between  $M_1$  and  $M_2$ . Do this and report the  $p$ -value for the test. Does  $M_2$  oversimplify the data compared with  $M_1$ ?

### Exercise 6.11

Use your knowledge of the English lexicon to explain why, in Table 6.4, [ts] and [sr] are so much more probable in the unigram model than in the other models, [st] is so much more probable in the bigram model than in the other models, and [tr] is so much less probable in the positional unigram model than in the other models.

### Exercise 6.12

In Section ??, the log-linear model of English onsets didn’t include any conditioning information  $X$ . What conditioning information  $X$  might you include in a richer model of English onset phonotactics?

### Exercise 6.13

Of the phonotactic models introduced in Section 6.10, which is the best predictive model with respect to the distribution of English onsets (as opposed to prediction of native speaker non-word acceptability judgments)? Assess the cross-validated log-likelihood achieved by each model using ten-fold cross-validation.

### Exercise 6.14

Consider the following frequency counts for the part of speech of the first word in each sentence of the parsed Brown corpus:



|                               |      |
|-------------------------------|------|
| (Pro)noun                     | 9192 |
| Verb                          | 904  |
| Coordinator                   | 1199 |
| Number                        | 237  |
| (Pre-)Determiner              | 3427 |
| Adverb                        | 1846 |
| Preposition or Complementizer | 2418 |
| <i>wh</i> -word               | 658  |
| Adjective                     | 433  |

Using a log-linear model with exactly one indicator feature function for each part of speech, demonstrate for yourself that the maximum-likelihood predictive distribution is simply the relative frequency estimate. Then introduce a Gaussian prior to your model. Plot the KL divergence from the MAP-estimated predictive distribution to the maximum-likelihood distribution as a function of the standard deviation of the prior.

### Exercise 6.15

How would you obtain confidence regions for parameter estimates in a Bayesian log-linear model? After reading Appendix ??, define and implement a Metropolis algorithm for sampling from the posterior distribution of a log-linear model with a Gaussian prior on the parameter estimates. Use this algorithm to generate confidence intervals for the feature weights in the models of Section 6.10.1. Which feature weights does the model have the most certainty about, and how should these features be interpreted? [HINT: you will save a lot of time if you use standard gradient-descent software to find the MAP-estimated feature weights and use these weights to initialize your sampling algorithm.]

### Exercise 6.16

The file `word_suffixes` contains frequency counts from CELEX (Baayen et al., 1995) for all suffixes of English word lemmas constructible from 17 English phonemes which are of length 2 or less.

- Define a small set of feature functions (no more than in the neighborhood of 20) on the basis of generalizations you see in the data and write a script to automatically extract the outputs of these feature functions for each form in the frequency database.
- Fit a maximum-likelihood log-linear model from this output. Inspect the feature weights and the predictive distribution. What conclusions can you draw from the results?
- Introduce a Gaussian prior and fit a new log-linear model using MAP estimation. How do the feature weights and the predictive distribution change?
- Based on any limitations you may see from the results of your first model, add new feature functions, refit the model, and discuss the changes you see between the simpler and more complex models.

**Exercise 6.17**

Use Bayes' Rule to show that when a fitted Poisson distribution with parameters  $\lambda$  is used to compute the probability distribution over counts in each response class subject to the constraint that the total count over all response classes is equal to 1, the resulting distribution is equivalent to that obtained by an unconditional log-linear model with the same parameters (see Figure 6.17).

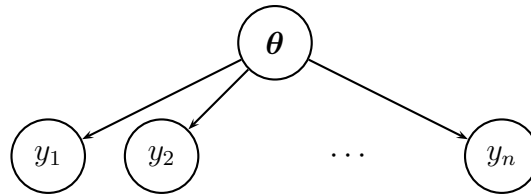
# Chapter 8

## Hierarchical Models

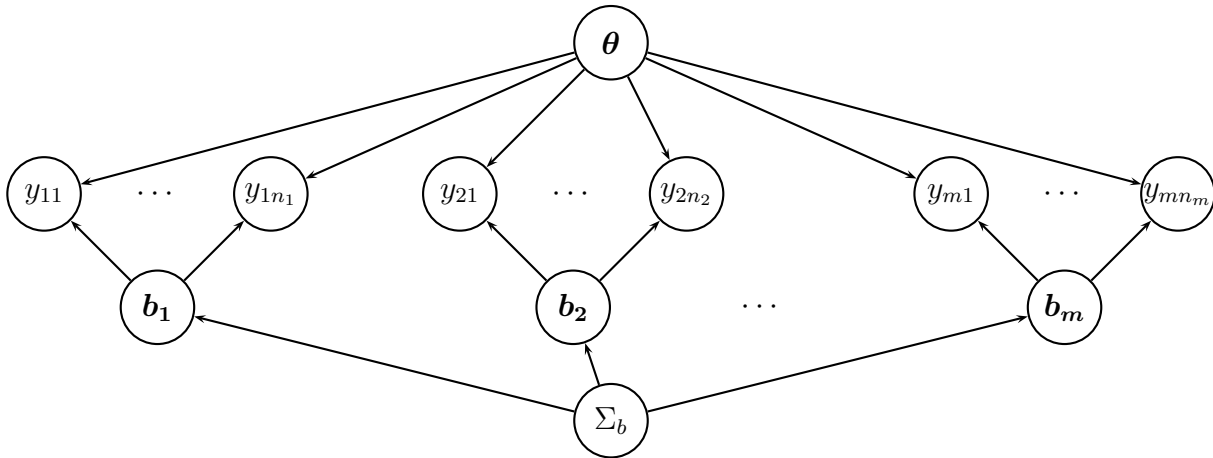
In the (generalized) linear models we've looked at so far, we've assumed that the observations are independent of each other given the predictor variables. However, there are many situations in which that type of independence does not hold. One major type of situation violating these independence assumptions is **CLUSTER-LEVEL ATTRIBUTES**: when observations belong to different clusters and each cluster has its own properties (different response mean, different sensitivity to each predictor). We'll now cover **HIERARCHICAL** (also called **MULTI-LEVEL** and, in some cases **MIXED-EFFECTS**) models, which are designed to handle this type of mutual dependence among datapoints. Common instances in which hierarchical models can be used include:

- Observations related to linguistic behavior are clustered at the level of the speaker, and speaker-specific attributes might include different baseline reading rates, differential sensitive to construction difficulty, or preference for one construction over another;
- Different sentences or even words may have idiosyncratic differences in their ease of understanding or production, and while we may not be able to model these differences, we may be able model the fact that there is incidental variation at the sentence or word level;
- Education-related observations (e.g., vocabulary size) of students have multiple levels of clustering: multiple measurements may be taken from a given student, multiple students may be observed from a class taught by a given teacher, multiple teachers may teach at the same school, multiple schools may be in the same city, and so forth.

This chapter introduces hierarchical models, building on the mathematical tools you have acquired throughout the book. This chapter makes considerably heavier use of Bayesian-style thinking and techniques than the previous chapter; this would be a good time to review marginalization (Section 3.2), Bayesian prediction and parameter estimation (Section 4.4), approximate posterior inference (Section 4.5), and confidence intervals (Chapter 5).



(a) A non-hierarchical model



(b) A simple hierarchical model, in which observations are grouped into  $m$  clusters

Figure 8.1: Non-hierarchical and hierarchical models

## 8.1 Introduction

The core idea behind the hierarchical model is illustrated in Figure 8.1. Figure 8.1a depicts the type of probabilistic model that we have spent most of our time with thus far: a model family has parameters  $\theta$ , which determine a probability distribution over outcomes, and a set of observations  $\mathbf{y}$  arises as a collection of independent draws from this distribution. Figure 8.1b illustrates the simplest type of hierarchical model: observations fall into a number of CLUSTERS, and the distribution over outcomes is determined jointly by (i) parameters  $\theta$  shared across clusters, and (ii) parameters  $\mathbf{b}$  which are shared among observations within a cluster, but may be different across clusters. Crucially, there is a second probability distribution, parameterized by  $\Sigma_{\mathbf{b}}$ , over the cluster-specific parameters  $\mathbf{b}$  themselves. In Figure 8.1b, there are  $m$  clusters, and for each cluster  $i$  there have been  $n_i$  observations  $y_{i1}, \dots, y_{in_i}$  made. All else being equal, we can expect that observations within a single cluster will tend to look more like each other than like observations in other clusters.

Let us consider a simple case in which the distribution from which the observations  $y_{ij}$  are drawn is characterized by just a few parameters—for example, they may be normally distributed, in which case the parameters are the mean and the variance. One natural type of clustering would be for each cluster to have its own mean  $\mu_i$  but for all the clusters to have the same variance  $\sigma_y^2$ . In the terminology of Figure 8.1b, we would classify the  $\mu_i$  as

cluster-specific parameters (in the  $\mathbf{b}$  nodes) and the variance  $\sigma_y^2$  as a shared parameter (in the  $\boldsymbol{\theta}$  node). In order to complete this probabilistic model, we would need to specify the distribution over the cluster-specific  $\mu_i$ . We might make this distribution normal as well, which requires two parameters of its own: a global mean  $\mu$  and variance  $\sigma_b^2$  (these would live in the  $\Sigma_b$  node). We can write this specification compactly as follows:

$$\begin{aligned}\mu_i &\sim \mathcal{N}(\mu, \sigma_b^2) \\ y_{ij} &\sim \mathcal{N}(\mu_i, \sigma_y^2)\end{aligned}\tag{8.1}$$

Equivalently, we could consider the cluster-specific parameters to be *deviations* from the overall mean  $\mu$ . In this approach, we would consider  $\mu$  as a shared  $\boldsymbol{\theta}$  parameter, and the mean of the deviations would be 0. We can compactly specify this version of the model as:

$$\begin{aligned}b_i &\sim \mathcal{N}(0, \sigma_b^2) \\ \mu_i &= \mu + b_i \\ y_{ij} &\sim \mathcal{N}(\mu_i, \sigma_y^2)\end{aligned}\tag{8.2}$$

The specifications in Equations 8.1 and 8.2 describe exactly the same family of probabilistic models. The advantage of the former specification is that it is more compact. The advantage of the latter specification is that the cluster-specific parameters are more directly interpretable as deviations “above” or “below” the overall average  $\mu$ . In addition, the latter specification leads to a nice connection with our discussion of linear models in Section 6.2. We can describe the same model as follows:

$$y_{ij} = \mu + \underbrace{b_i}_{\sim \mathcal{N}(0, \sigma_b^2)} + \underbrace{\epsilon_{ij}}_{\sim \mathcal{N}(0, \sigma_y^2)}\tag{8.3}$$

That is, an individual observation  $y_{ij}$  is the sum of the overall mean  $\mu$ , a normally-distributed cluster-level deviation  $b_i$ , and a normally-distributed observation-level deviation  $\epsilon_{ij}$ .

Let us now consider a concrete example with slightly greater complexity. Suppose that a phonetician is interested in studying the distribution of the pronunciation of the vowel [a], recruits six native speakers of American English, and records each speaker once a week for fifteen weeks. In each case, the phonetician computes and records the F1 and F2 formants of the pronounced syllable. Now, no two recordings will be exactly alike, but different individuals will tend to pronounce the syllable in different ways—that is, there is both within-individual and between-individual variation in F1 formant from recording to recording. Let us assume that inter-speaker (cluster-level) variation and inter-trial (observation-level) variation are both multivariate-normal. If we denote the  $\langle \text{F1}, \text{F2} \rangle$  value for the  $j$ th recording of speaker  $i$  as  $y_{ij}$ , then we could write our model as follows:

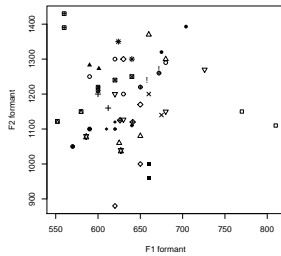


Figure 8.2: Empirically observed male adult speaker means for first and second formants of [ɑ]

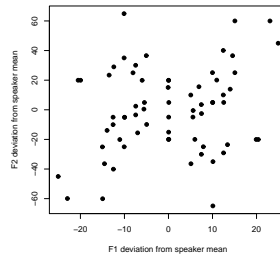


Figure 8.3: Empirically observed deviations from speaker means

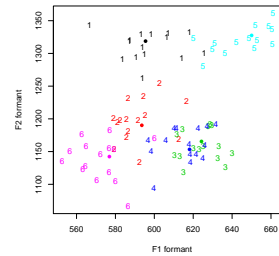


Figure 8.4: Simulated formant data for six speakers. Speaker-specific means  $\mu + b_i$  are given as filled circles.

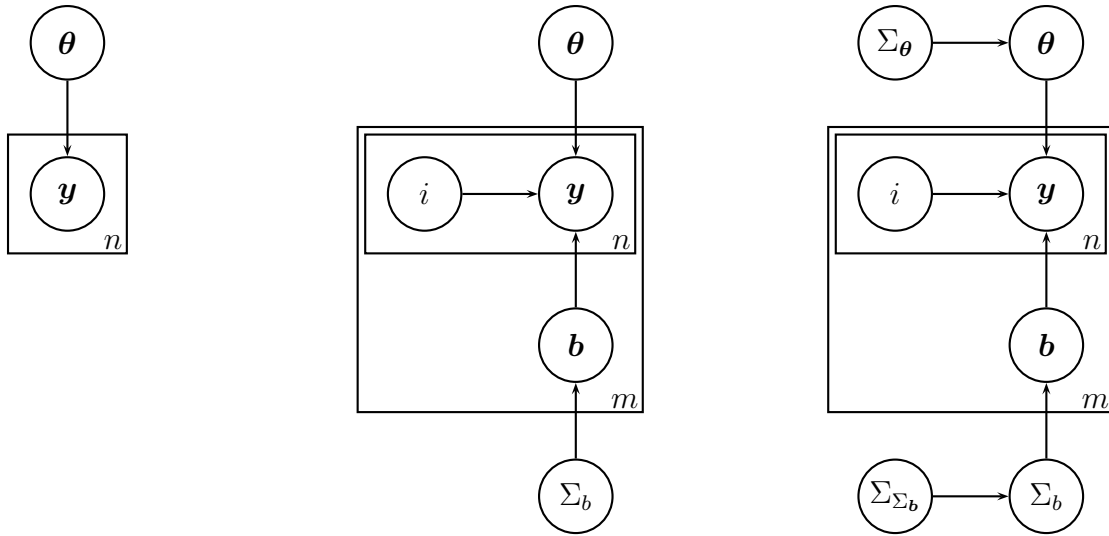
$$\begin{aligned}
 b_i &\sim \mathcal{N}(\mathbf{0}, \Sigma_b) \\
 \mu_i &= \mu + b_i \\
 y_{ij} &\sim \mathcal{N}(\mu_i, \Sigma_y)
 \end{aligned}
 \tag{8.4}$$

where  $\mathbf{0}$  is the vector  $\langle 0, 0 \rangle$ .

The only difference between the models in Equations 8.2 and 8.4 is that whereas the former is univariate, the latter is multivariate:  $b_i$  is distributed around zero according to some covariance matrix  $\Sigma_b$ , and the  $y_{ij}$  are distributed around  $\mu_i$  according to another covariance matrix  $\Sigma_y$ . Both the univariate and multivariate models (Equations 8.2 and 8.4) have precisely the structure of Figure 8.1b, with  $\mu$  and  $\Sigma_y$  being the shared parameters  $\theta$ . For the moment, to estimate the model parameters we use the simple expedient of using the sample mean and covariance of speaker-averages of adult-male data due to Peterson and Barney (1952) (shown in Figure 8.2) to estimate  $\mu$ , and  $\Sigma_b$ , and the covariance of deviations from speaker means (shown in Figure 8.3) to estimate  $\Sigma_y$ . Figure 8.4 gives sample data generated from this model. The individual speakers correspond to clusters of trial-level observations. Note how there is considerable intra-cluster variation but the variation between clusters is at least as large.

## 8.2 Parameter estimation in hierarchical models

The previous section outlines what is essentially the complete probabilistic theory of hierarchical models. However, the problems of statistical inference within hierarchical models require more discussion. Before we dive into these issues, however, it is worthwhile to introduce a more succinct graphical representation of hierarchical models than that used in Figure 8.1b. Figure 8.5a is a representation of non-hierarchical models, as in Figure 8.1a, where the individual observations  $y_i$  have been collapsed into a single node  $\mathbf{y}$ . The box



(a) The non-hierarchical model of Figure 8.1a    (b) The simple hierarchical model of Figure 8.1b    (c) The model of Figure 8.1b with priors over  $\theta$  and  $\Sigma_b$

Figure 8.5: A more succinct representation of the models in Figure 8.1

labeled “ $n$ ” surrounding the  $y$  node indicates that  $n$  independent events are generated at this node; the labeled box is called a PLATE. Likewise, Figure 8.5b is a representation of the class of simple hierarchical models shown in Figure 8.1b, with both individual observations  $y_{ij}$  and class-specific parameters  $b_i$  compressed into single nodes. The outer plate indicates that  $m$  independent events are generated at the  $b$  node; the inner plate (embedded in the outer plate) indicates that for the  $i$ -th of these  $m$  events,  $n_i$  sub-events are generated. Each sub-event has a “cluster identity” label—the node labeled  $i$ , which allows us to track which cluster each observation falls into—and the nodes  $b$ ,  $\theta$ , and  $i$  jointly determine the distribution over the outcome at the  $y$  node for this sub-event.

In light of this picture, let us consider the problem of parameter estimation for a case such as the formant measurements of the previous section. We know our observations  $y$ , and we also know the cluster identity variable  $i$ —that is, which individual produced each observation. We do not know the shared model parameters  $\theta$  the parameters  $\Sigma_b$  governing cross-speaker variability, or the speaker-specific variations  $b_i$  themselves. Upon reflection, however, it should become clear that the primary goal of the study is to learn  $\theta$  and  $\Sigma_b$ , not  $b_i$ . Yet the  $b_i$  stand in the way of estimating  $\Sigma_b$ . It might seem like a good idea to first construct point estimates of  $b_i$  and then use these estimates directly to estimate  $\Sigma_b$ , but this approach throws away valuable information (our uncertainty about the true values of the  $b_i$ , which we should take into account). How can we make inferences about our parameters of interest in a principled way?

The answer is actually quite simple: whatever technique of parameter estimation we choose, we should *marginalize* over the cluster-specific  $b_i$ . This leads us to two basic approaches to parameter estimation for hierarchical models:

1. Construct point estimates of the parameters of interest,  $\widehat{\Sigma}_{\mathbf{b}}$  and  $\widehat{\boldsymbol{\theta}}$ , using the principle of maximum likelihood (or a close variant). The likelihood to be maximized is marginal over  $\mathbf{b}$ :

$$\text{Lik}(\Sigma_{\mathbf{b}}, \boldsymbol{\theta}; \mathbf{y}) = \int_{\mathbf{b}} P(\mathbf{y}|\boldsymbol{\theta}, \mathbf{b})P(\mathbf{b}|\Sigma_{\mathbf{b}}, i) d\mathbf{b} \quad (8.5)$$

2. Use Bayesian inference: introduce prior distributions over  $\boldsymbol{\theta}$  and  $\Sigma_{\mathbf{b}}$ , and compute the (approximate) posterior distribution over  $\boldsymbol{\theta}$  and  $\Sigma_{\mathbf{b}}$ . Since the introduction of priors means that  $\boldsymbol{\theta}$  and  $\Sigma_{\mathbf{b}}$  are themselves drawn from some distribution, this is actually a shift to a slightly more complex hierarchical model, shown in Figure 8.5c.  $\Sigma_{\boldsymbol{\theta}}$  and  $\Sigma_{\Sigma_{\mathbf{b}}}$ , chosen by the researcher, parameterize the priors over  $\boldsymbol{\theta}$  and  $\Sigma_{\mathbf{b}}$  respectively. Via Bayes' rule, the posterior distributions of interest can be written as follows:

$$P(\Sigma_{\mathbf{b}}, \boldsymbol{\theta}|\mathbf{y}) \propto \int_{\mathbf{b}} P(\mathbf{y}|\boldsymbol{\theta}, \mathbf{b})P(\mathbf{b}|\Sigma_{\mathbf{b}}, i)P(\boldsymbol{\theta}|\Sigma_{\boldsymbol{\theta}})P(\Sigma_{\mathbf{b}}|\Sigma_{\Sigma_{\mathbf{b}}}) d\mathbf{b} \quad (8.6)$$

Note that the posterior distribution looks very similar to the unnormalized likelihood of Equation (8.5) above. The only difference is, as always, the presence of the prior probability  $P(\boldsymbol{\theta}|\Sigma_{\boldsymbol{\theta}})P(\Sigma_{\mathbf{b}}|\Sigma_{\Sigma_{\mathbf{b}}})$  in the posterior distribution. It's worth recalling at this point that if the prior distribution is chosen to be uniform, then the maximum-likelihood estimate is also the Bayesian maximum a-posteriori (MAP) estimate.

We'll now illustrate each of these approaches with reference to actual formant data from recordings of adult male American speakers' pronunciations of [a] by Peterson and Barney (1952). The F1 data are plotted in Figure 8.6.

### 8.2.1 Point estimation based on maximum likelihood

We illustrate the point estimation approach by treating the F1 and F2 formants separately.<sup>1</sup> For each of the formants, we assume the model of Equation (8.2): that the speaker-specific deviations from the grand mean are normally distributed, and that trial-specific deviations are also normally distributed. We reiterate that there are three parameters to be estimated in each model: the overall mean  $\mu$ , the inter-speaker variance  $\Sigma_{\mathbf{b}}$ , and the intra-speaker, inter-trial variance  $\Sigma_{\mathbf{y}}$ . The maximum-likelihood estimates for these parameters can be seen in the output of each model fit:

|                       | F1    | F2     |
|-----------------------|-------|--------|
| $\mu$                 | 630.6 | 1191.9 |
| $\sigma_{\mathbf{b}}$ | 43.1  | 100.8  |
| $\sigma_{\mathbf{y}}$ | 16.9  | 40.1   |

---

<sup>1</sup>The R package lme4 is the state of the art in likelihood-based point estimation for a wide variety of hierarchical models, and we use it here.



In this case, there is considerably more inter-speaker variation than intra-speaker variation. Eyeballing Figure 8.6 and comparing it with the parameter estimates above, we can see that the overall mean  $\mu$  is right at the grand mean of all the observations (this has to occur because the dataset is balanced in terms of cluster size), and that nearly all of the observations lie within two cluster-level standard deviations (i.e.  $\sigma_{\mathbf{b}}$ ) of the grand mean.

### Conditional estimates of cluster-specific parameters

In the point-estimation approach, we have focused on the parameters of interest— $\boldsymbol{\theta}$  and  $\Sigma_{\mathbf{b}}$ —while maintaining our ignorance about  $\mathbf{b}$  by marginalizing over it. Nevertheless, in many cases we may be interested in recovering information about  $\mathbf{b}$  from our model. For example, although the ultimate point of the phonetic study above was to estimate inter-speaker and intra-speaker variation in the pronunciation of [ɑ], we might also be peripherally interested in making inferences about the average pronunciation behavior of the specific individuals who participated in our study. Formally, the point estimates of  $\boldsymbol{\theta}$  and  $\Sigma_{\mathbf{b}}$  determine a conditional probability distribution over  $\mathbf{b}$ . The mode of this distribution is called the BEST LINEAR UNBIASED PREDICTOR (BLUP)  $\hat{\mathbf{b}}$ :

$$\hat{\mathbf{b}} \stackrel{\text{def}}{=} \arg \max_{\mathbf{b}} P(\mathbf{b} | \hat{\boldsymbol{\theta}}, \hat{\Sigma}_{\mathbf{b}}, \mathbf{y})$$

The F1 BLUPs for speakers in the current example are plotted as magenta circles in Figure 8.7.

### Shrinkage

Another way of estimating speaker-specific averages would simply be to take mean recorded F1 frequency for each speaker. But these two approaches lead to different inferences. Figure 8.7 shows the deviation of each speaker’s mean recorded F1 frequency from the grand mean as black squares; recall that the conditional estimate  $\hat{\mathbf{b}}$  are magenta circles. Notice that the conditional modes are systematically closer to zero than the means of the raw trials; this effect is more dramatic for speakers with larger deviations. This happens because the finite variance of  $\mathbf{b}$  in the hierarchical model penalizes large deviations from the grand mean  $\mu$ . This effect is called SHRINKAGE and is ubiquitous in hierarchical models.

## 8.2.2 Bayesian posterior inference in hierarchical models

We can compare the point estimates (with standard errors) that we obtained in Section 8.2.1 with posterior estimates obtained using Bayesian inference. We specify the hierarchical model as follows:

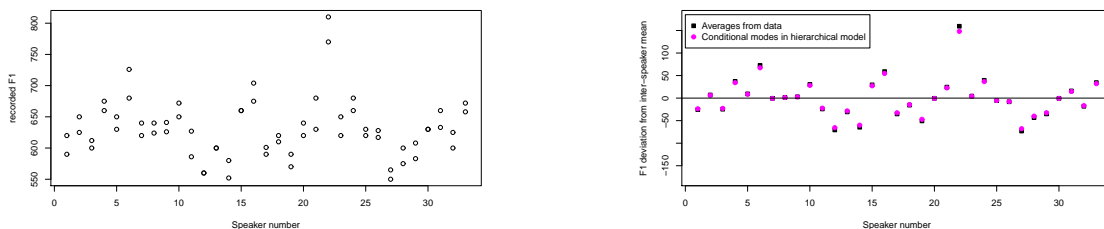


Figure 8.6: Observed F1 measurements by Speaker number      Figure 8.7: Conditional estimates of speaker-specific mean F1 frequencies, and shrinkage

$$\begin{aligned}
 \mu &\sim \mathcal{N}(0, 10^5) \\
 \log \sigma_{\mathbf{b}} &\sim \mathcal{U}(-100, 100) \\
 \log \sigma_{\mathbf{y}} &\sim \mathcal{U}(-100, 100) \\
 b_i &\sim \mathcal{N}(0, \sigma_{\mathbf{b}}^2) \\
 \mu_i &= \mu + b_i \\
 y_{ij} &\sim \mathcal{N}(\mu_i, \sigma_{\mathbf{y}}^2)
 \end{aligned} \tag{8.7}$$

This is exactly the same model as in Equation 8.2, with the addition of three extra lines at the top specifying prior distributions over the overall mean  $\mu$ , inter-speaker variability  $\sigma_{\mathbf{b}}$ , and intra-speaker variability  $\sigma_{\mathbf{y}}$ . There are two important points regarding this model specification. The first is that we are using a normal distribution with very large variance as a prior on the grand mean  $\mu$ . The normal distribution is conjugate (Section 4.4.3) to the mean of a normal distribution, which has computational advantages; the large variance means that the prior is relatively uninformative, placing little constraint on our inferences about likely values of  $\mu$ . The second point is how we are defining the priors on the variance parameters  $\sigma_{\mathbf{b}}$  and  $\sigma_{\mathbf{y}}$ . Although the inverse chi-squared distribution (Section B.4) is conjugate to the variance parameter of a normal distribution, this distribution does not lend itself well to an uninformative specification. As described earlier in Section 4.5, placing a uniform distribution over the log of the standard deviation allows our prior to be uninformative in a “scale-free” sense.<sup>2</sup>

Using the sampling techniques via BUGS described in Section 4.5, we can obtain approximate highest-posterior density confidence intervals and conditional modes (the latter being approximated by a very narrow HPD interval), and plot estimates of the posterior (Figure 8.8). For F1 formants, these 95% HPD intervals and posterior modes are:

The posterior simulations are in broad agreement with the point estimates and standard errors obtained in Section 8.2.1. We leave obtaining similar results for F2 as Exercise 8.3.

---

<sup>2</sup>Good discussion of practical choices for priors on variance parameters can be found in Gelman et al. (2004, Appendix C).

|                       | lower bound | upper bound | posterior mode |
|-----------------------|-------------|-------------|----------------|
| $\mu$                 | 613.6       | 644.7       | 618.9          |
| $\sigma_{\mathbf{b}}$ | 35.4        | 60.4        | 43.7           |
| $\sigma_{\mathbf{y}}$ | 13.5        | 20.9        | 19.1           |

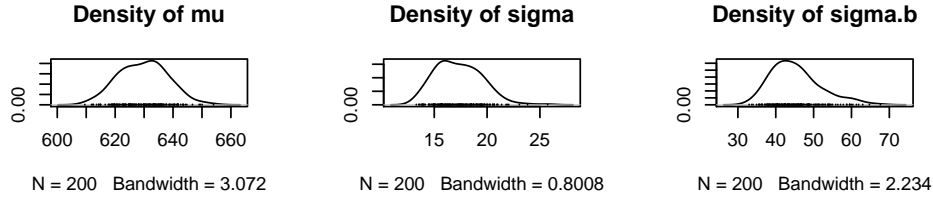


Figure 8.8: Output of MCMC sampling on hierarchical model parameters

### 8.2.3 Multivariate responses

One shortcoming of the analysis of the previous two sections is that F1 and F2 formants were analyzed separately. Correlations between F1 and F2 frequency are captured at neither the inter-speaker nor the intra-speaker level. However, there is a hint of such a correlation in the Figure 8.3. This raises the question of whether such a correlation is reliable at either level. We can address this question directly within a hierarchical model by using bivariate representations of  $\mathbf{y}$  and  $\mathbf{b}$ . We'll illustrate this type of analysis in a Bayesian framework. The model specification looks similar to the univariate case given in Equation 8.7, but we are using different prior distributions because our normal distributions of interest are multivariate (even though they are still written as  $\mathcal{N}$ ):

$$\begin{aligned}
 b_i &\sim \mathcal{N}(\mathbf{0}, \Sigma_{\mathbf{b}}) \\
 \mu_i &= \mu + b_i \\
 y_{ij} &\sim \mathcal{N}(\mu_i, \Sigma_{\mathbf{y}}) \\
 \mu &\sim \mathcal{N}(\mathbf{0}, \begin{bmatrix} 10^5 & 0 \\ 0 & 10^5 \end{bmatrix}) \\
 \Sigma_{\mathbf{b}} &\sim \mathcal{IW} \left( \begin{bmatrix} 10^5 & 0 \\ 0 & 10^5 \end{bmatrix}, 2 \right) \\
 \Sigma_{\mathbf{y}} &\sim \mathcal{IW} \left( \begin{bmatrix} 10^5 & 0 \\ 0 & 10^5 \end{bmatrix}, 2 \right)
 \end{aligned}$$

The symbol  $\mathcal{IW}$  stands for the inverse Wishart distribution, which is a widely-used prior distribution for covariance matrices; the large diagonal entries in the matrix parameterizing it signal uninformativity, and the zero off-diagonal entries signal that there is no particular prior expectation towards correlation between F1 and F2 deviations. The inverse Wishart distribution is described more completely in Section B.7.

There are eight distinct parameters in our model over which we would like to make inferences: two  $\mu$  parameters, three  $\Sigma_{\mathbf{b}}$  parameters, and three  $\Sigma_{\mathbf{y}}$  parameters (recall from

Section 3.5 that  $\Sigma_{\mathbf{b}}$  and  $\Sigma_{\mathbf{y}}$  have the form  $[\begin{smallmatrix} \sigma_{21}^{11} & \sigma_{22}^{12} \\ \sigma_{21}^{12} & \sigma_{22}^{22} \end{smallmatrix}]$  but that  $\sigma_{12} = \sigma_{21}$ ). Using BUGS once more we obtain the following 95% HPD confidence intervals from the posterior distribution:

|                       | lower bound   | higher bound   | posterior mode   |
|-----------------------|---|--|--|
| $\mu$                 | $\langle 486.4, 1000.3 \rangle$                             | $\langle 499.3, 1157 \rangle$                                | $\langle 490.2, 1120.8 \rangle$                              |
| $\Sigma_{\mathbf{b}}$ | $\begin{bmatrix} 6.3 & -1 \\ -1 & 78.4 \end{bmatrix}$       | $\begin{bmatrix} 11 & -1 \\ -1 & 128.7 \end{bmatrix}$        | $\begin{bmatrix} 8.7 & -1 \\ -1 & 117.4 \end{bmatrix}$       |
| $\Sigma_{\mathbf{y}}$ | $\begin{bmatrix} 120.3 & 0.81 \\ 0.81 & 61.4 \end{bmatrix}$ | $\begin{bmatrix} 175.7 & 0.99 \\ 0.99 & 209.9 \end{bmatrix}$ | $\begin{bmatrix} 144.4 & 0.96 \\ 0.96 & 136.3 \end{bmatrix}$ |

Of particular interest are the inferences about correlations between F1 and F2 formants, which are the most compelling reason to do multivariate analysis in the first place. In the above analysis, the posterior mode suggests a negative F1-F2 correlation at the inter-speaker level, but positive correlation at the intra-speaker level. However, the confidence intervals on these correlations show that this suggestion is far from conclusive.

It is instructive to compare this analysis to a more straightforward analysis of F1-F2 correlation in which inter-speaker correlation is estimated by calculating speaker means and computing a correlation coefficient on these means, and intra-speaker correlation is estimated obtained by simply subtracting out the speaker-specific mean from each observation and then calculating a correlation coefficient on the resulting residuals. For inter-speaker variation, it gives an empirical correlation coefficient of  $r = -0.01$ , with a 95% confidence interval of  $[-0.35, 0.34]$ ; for intra-speaker variation, it gives an empirical correlation coefficient of  $r = 0.24$ , with a 95% confidence interval of  $[-0.002, 0.46]$ .<sup>3</sup> Although this approach also leads to the conclusion that there is no reliable F1-F2 correlation at either the inter-speaker or intra-speaker level, these confidence intervals indicate considerably more certainty in the true correlation than the Bayesian HPD intervals suggest, and the  $p$ -value for correlation at the intra-speaker level is very close to significant at 0.052. A key point here is that this latter approach based on empirical speaker means doesn't take into account the uncertainty about true speaker means, and thus leads to conclusions about inter-speaker variation of greater certainty than may be warranted. The Bayesian HPD interval takes this uncertainty into account, and is thus more agnostic about the true correlation.

### 8.3 Hierarchical linear models

Now we'll move on from hierarchical models of the form in Equation 8.3 to conditional models. Intuitively, this means we'll be estimating distributions of the form

$$P(Y|X, i) \tag{8.8}$$

---

<sup>3</sup>This confidence interval is derived by using the transform  $z = \frac{1}{2} \log \frac{1+r}{1-r}$ ; the resulting  $z$  is approximately normally distributed (Cohen et al., 2003, p. 45), and so a confidence interval based on the normal distribution can be calculated as described in Section 5.3.

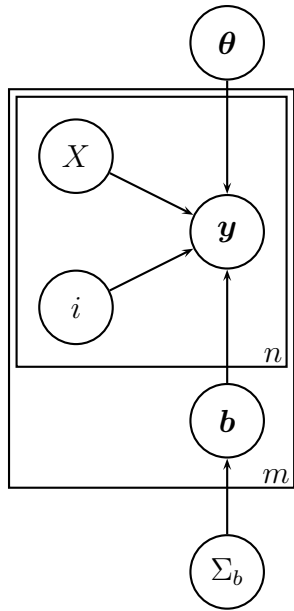


Figure 8.9: A conditional hierarchical model for probability distributions of the form  $P(Y|X, i)$

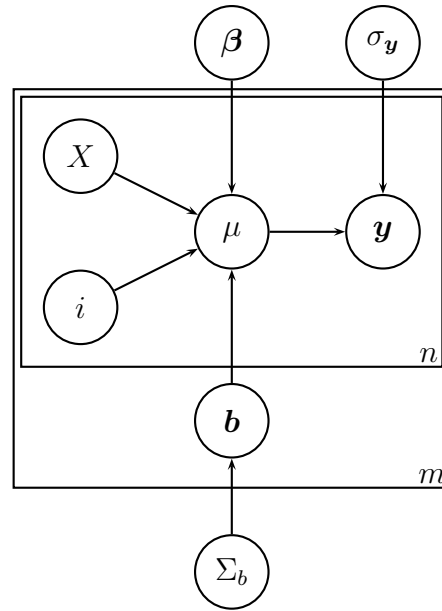


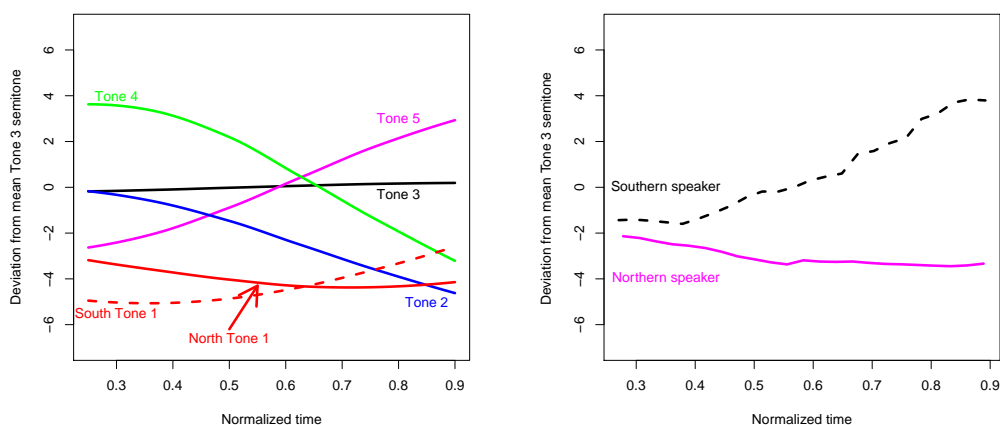
Figure 8.10: A structure more specific to hierarchical linear models, where the influence of cluster identity  $i$  and covariates  $X$  on the outcome  $y$  is only on the predicted mean  $\mu$ .

where  $X$  are the covariates (there can be many of them) and  $i$  are the cluster identities. Figure 8.9 illustrates this type of model. The only change from Figure 8.5b is the addition of the covariates  $X$  as a separate node in the graph. Once again, the primary targets of inference are typically  $\theta$  and  $\Sigma$ , and we'd want to marginalize over  $\mathbf{b}$  in making our inferences about them.

We'll start with a study of hierarchical linear models. Assume that we have covariates  $X_1, \dots, X_M$  on which we want to condition  $Y$ . We can express the  $j$ -th outcome in the  $i$ -th cluster as

$$y_{ij} = \alpha + b_{i0} + (\beta_1 + b_{i1})X_1 + \dots + (\beta_M + b_{iM})X_M + \epsilon \quad (8.9)$$

where  $\epsilon$  is, as before, normally distributed. This equation means that every cluster  $i$  has a cluster-specific intercept  $\alpha + b_{i0}$  and a slope parameter  $\beta_k + b_{ik}$  that determines the contribution of covariate  $X_k$  to the mean outcome. In the notation of Figure 8.9, the parameters  $\alpha$  and  $\{\beta_k\}$ , along with the variability  $\sigma_y$  governing  $\epsilon$ , are  $\theta$  parameters, shared across clusters, whereas the  $b_{ik}$  parameters are specific to cluster  $i$ . Figure 8.10 shows a slightly more nuanced picture illustrating how the predicted mean mediates the influence of covariates and cluster identity on the outcome; here, only  $\alpha$  and  $\{\beta_k\}$  are  $\beta$  parameters. Equation 8.9 describes the most general case, where all predictors have both shared parameters and cluster-specific



(a) Mean Sui tone contours

(b) Two specific trial Sui tone 1 contours for the word *jun*

Figure 8.11: Sui tone contours

parameters. However, the models can be constrained such that some predictors have only shared parameters and some others have only cluster-specific parameters.

### Fitting and drawing inferences from a hierarchical linear model: practice

We'll illustrate the utility of hierarchical linear models with a simple instance in which the covariates are categorical. Stanford (2008) investigated variability in the low lexical tone contour of Sui, a minority language in Southwest China. The Sui practice clan exogamy: a wife and husband must originate from different clans, and the wife immigrates to the husband's village. Both Northern and Southern Sui clan dialects have six tones, but they have different low-tone (Tone 1) pitch contours. Figure 8.11 illustrates the mean contours for five of the six Sui tones, along with sample tone contours taken from individual recordings of one southern Sui and one northern Sui speaker (who lived in their home village). According to Stanford (2008), the difference in this tone contour is audible but subtle, and the Sui do not mention it as a hallmark of the tone differences between northern and southern clan dialects. Stanford investigated two questions:

1. whether this tone contour can be reliably measured; and
2. whether immigrant Sui speakers adopt the lexical tone contour of their husband's clan, or keep their original tone contour.

We begin with question 1. Stanford observed that in tone 1, from the temporal midway point of each tone to the end, the mean tone contour is fairly straight, but it tends to rise for Southern speakers whereas it stays flat for Northern speakers (Figure 8.11a). Therefore one difference between northern and southern tone 1 contour may be characterizable by the *slope*

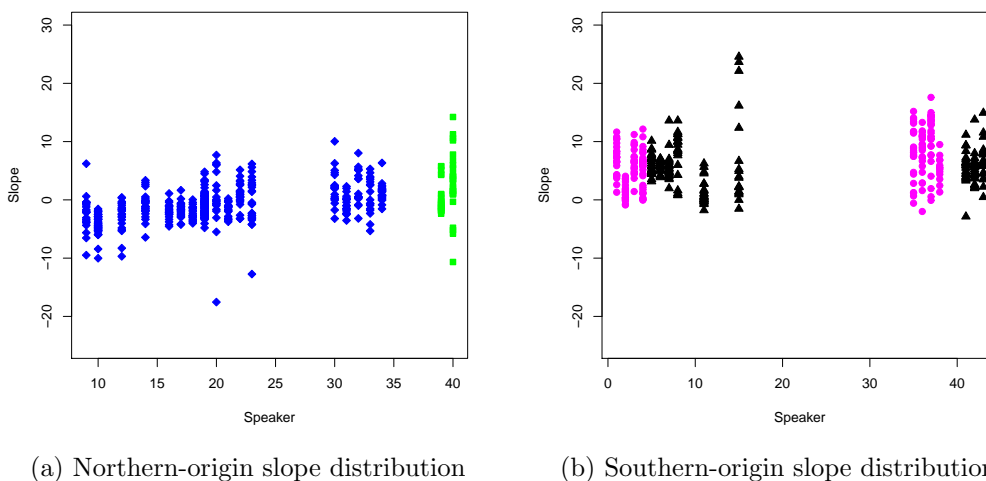


Figure 8.12: Distribution of slopes for speakers originating northern and southern clans. Southward and northward migrants are plotted in green squares and magenta circles respectively.

(in the sense of linear regression) of the tone contour in this part of the syllable. Figure 8.12 plots the distribution of tone contour slopes for each individual trial for Northern-origin and Southern-origin speakers. There is an apparent trend for Northern speakers to have lower slopes. However, there is also an apparent trend for different speakers of each origin to have idiosyncratically different slopes. We could deal with this nesting structure through analysis of variance with speaker as a random factor (Section 6.6), but the data are unbalanced, which is not ideal for analysis of variance. Lack of balance presents no fundamental difficulty for a hierarchical linear model, however.

### 8.3.1 Hypothesis testing in hierarchical linear models

We begin with a question that is something of an aside: how can we test for significant inter-speaker variability in mean tone contour slope? This can be done in a frequentist paradigm, via model comparison between models fitted using maximum likelihood, with and without inter-speaker variability, using the likelihood ratio test (Section 5.4.4). The  $p$ -values obtained by the likelihood-ratio test for models differing in the number of parameters governing inter-cluster variability (“random-effects” structure, or  $\Sigma_{\mathbf{b}}$ ) are CONSERVATIVE (Stram and Lee, 1994; Pinheiro and Bates, 2000; Baayen et al., 2008), meaning that the true  $p$ -value will generally be smaller (i.e. more significant) than the  $p$ -value obtained by consulting the  $\chi^2$  distribution. Conservativity is good in the sense that an obtained significant  $p$ -value can be trusted, but dangerous in the sense that a large (i.e. insignificant)  $p$ -value is not necessarily grounds to exclude inter-cluster variability from the model. The conventional wisdom in this regard is, when in doubt, to include all potentially relevant cluster-level parameters in one’s

model, unless doing so introduces computational problems.

To illustrate this type of hypothesis test on cluster-level parameters, we construct two models, a simpler model  $M_0$  with shared parameters for speaker origin ( $SO$ , valued 1 for southern origin and 0 for northern origin), migration north ( $MN$ ), and migration south ( $MS$ ), and no cluster-level parameters; and a more general model  $M_1$  with the same shared parameters as  $M_0$  but also a normally-distributed speaker-level slope deviation,  $\mathbf{b}$ , distributed as  $\mathcal{N}(0, \sigma_{\mathbf{b}})$ . In the style of Equation 8.3, we can write  $M_1$  as

$$y_{ij} = \alpha + \beta_1 SO + \beta_2 MN + \beta_3 MS + \underbrace{b_i}_{\sim \mathcal{N}(0, \sigma_b^2)} + \underbrace{\epsilon_{ij}}_{\sim \mathcal{N}(0, \sigma_y^2)}$$

and  $M_0$  looks the same except that the  $b_i$  term is absent. The data to which the model is fitted are shown in Figure 8.12. It turns out that the log-likelihood of the MLE fit of  $M_0$  is -2403.8 whereas the log-likelihood for  $M_1$  is -2308.7. Consulting twice the difference of these log-likelihoods against the  $\chi_1^2$  distribution, we find that the improvement of  $M_1$  over  $M_0$  is extremely unlikely under  $M_0$  ( $p \ll 0.001$ ). It's also worth looking at our estimates of inter-speaker and intra-speaker variability in this case. Consulting Table 8.1, we see that the MLE for  $\sigma_{\mathbf{b}}$  is 3.35, and for  $\sigma_y$  it is 3.07, so this case is unlike English F1 and F2 formants in that there is much more intraspeaker variability than interspeaker variability; this conclusion is broadly consistent with the visual picture in Figure 8.12. Nevertheless, there is extremely strong evidence that speakers vary idiosyncratically in their average tone 1 contours, above and beyond the variability across place of origin.

We now move to the question of hypothesis testing involving shared model parameters (“fixed effects” structure, or  $\theta$ ). Perhaps surprisingly, how to test for significance of an effect of a shared model parameter is a matter of some current controversy. In principle, one could use the likelihood ratio test to compare a more complex model with a simpler model. However, it has been argued that the likelihood ratio test will lead to ANTI-CONSERVATIVE  $p$ -values (i.e. the true  $p$ -value is less significant than the  $p$ -value obtained by consulting the  $\chi^2$  distribution) for comparison of models with the same cluster-level parameters but different shared parameters (Pinheiro and Bates, 2000, pp. 87–92).<sup>4</sup> This leaves two approaches currently in vogue. On the first approach, a single model is fitted with the method of maximum likelihood, and for the shared parameter of interest, the parameter estimate and its standard error are used to obtain a  $p$ -value based on the  $t$  statistic, just as in standard linear regression (Section 6.4; for testing multiple parameters simultaneously, an  $F$ -test is used). This approach itself carries a degree of controversy involving how many degrees of freedom the relevant  $t$  distribution should be assumed to have. As a rule of thumb, however, if there are many more observations than model parameters, the  $t$  distribution is generally taken to be approximated by the standard normal distribution (see also Section B.5). This is illustrated below; the MLEs, standard errors, and resulting  $t$  statistics for our four parameters

---

<sup>4</sup>As a caveat, it is not clear to this author that the anti-conservativity involved is appreciable unless the number of total parameters in the model is quite large relative to the number of observations, which quite often is *not* the case for linguistic datasets.



|                       | MLE $[\hat{\beta}]$ | Standard Error $[SE(\hat{\beta})]$ | $t$ statistic $\left[ \stackrel{\text{def}}{=} \frac{\hat{\beta}}{SE(\hat{\beta})} \right]$ |
|-----------------------|---------------------|------------------------------------|---|
| $\sigma_{\mathbf{b}}$ | 3.35                |                                    |   |
| $\sigma_{\mathbf{y}}$ | 3.07                |                                    |   |
| Intercept             | -0.72               | 0.47                               | -1.53   |
| $SO$                  | 7.21                | 0.83                               | 8.67  |
| $MN$                  | 2.38                | 1.44                               | 1.65  |
| $MS$                  | -0.82               | 0.94                               | -0.87   |

Table 8.1: Shared parameter estimates and standard errors for  $M_1$ . Note that standard errors are not appropriate for estimates of  $\sigma_{\mathbf{b}}$  or  $\sigma_{\mathbf{y}}$ , as these are not normally distributed.

of interest are given in Table 8.1.

The second approach currently in vogue is to use Bayesian inference and bypass the classical hypothesis testing paradigm altogether. Instead, one can estimate a Bayesian confidence region (Section 5.1) on the shared parameters  $\theta$ , by sampling from the posterior distribution over  $\theta$  using Markov Chain Monte Carlo (MCMC) sampling as discussed in Section 4.5, earlier in this chapter, and in Appendix ???. Here we illustrate this approach by sampling from the posterior in Figure 8.10 over  $P(\beta|\mathbf{y}, \widehat{\Sigma}_{\mathbf{b}}, \widehat{\sigma}_{\mathbf{y}})$ , performed by the `lme4` package and giving us the following 95% HPD confidence intervals and posterior modes:

|          | lower bound | upper bound | posterior mode |
|----------|-------------|-------------|----------------|
| $\alpha$ | -1.53       | 0.13        | -1.06          |
| $SO$     | 5.66        | 8.69        | 8.41           |
| $MN$     | -0.06       | 4.94        | 3.08           |
| $MS$     | -2.58       | 0.88        | -0.07          |

On the Bayesian interpretation, we can be over 95% certain that the true parameter estimate for the effect of being from the south (with respect to the reference level of clan origin, the north) is positive. It has recently become popular in the psycholinguistics literature to call the largest value  $q$  such that  $1 - \frac{q}{2}$  of the posterior probability mass on a parameter  $\theta$  lies on one side of zero a "MCMC-based  $p$ -value" for  $\theta$  (Baayen, 2008). Although this value  $q$  is certainly a useful heuristic for assessing the strength of the evidence supporting a meaningful role for  $\theta$  in the model, it is also worth keeping in mind that this value  $q$  is NOT a  $p$ -value in the traditional Neyman-Pearson paradigm sense of the term.

### Simultaneous assessment of significance of multiple parameters

We now turn to another question: whether migration into a new clan has a reliable effect on tone 1 slope. If it did, then the theoretically sensible prediction would be that migration of a southern-origin woman to the north should tend to lower the slope, and migration of a northern woman to the south should tend to raise the slope. To test this possibility, we can consult model  $M_1$ , whose parameters associated with variables  $MN$  and  $MS$  encode this effect. Looking at Table 8.1, we see that both coefficients associated with migration are consistent with the theoretical prediction. There is considerable uncertainty about each

parameter (as indicated by the  $t$ -value), but what we would really like to do is to assess the overall explanatory benefit accrued by introducing them together. The conceptually simplest way to do this is to use the  $F$  statistic from model comparison between  $M_1$  and a simpler model  $M'_0$  in which effects of  $MN$  and  $MS$  are absent. Computing an  $F$ -test between these two models we obtain an  $F$  statistic of 1.743. In order to evaluate statistical significance, we need to choose which  $F$  distribution should serve as the reference distribution, but as noted earlier, there is some controversy as to how many degrees of freedom to use in the *denominator* for such an  $F$  statistic. However, we can play the devil's advocate momentarily and ask how much evidence would exist for an effect of migration in the most optimistic interpretation. The maximum degrees of freedom for the  $F$  statistic denominator is the number of observations minus the number of parameters in the full model, or 887. The cumulative distribution function for  $F_{2,887}$  at 1.743 is 0.824, hence the best-case  $p$ -value is 0.176, and the overall effect is thus marginal at best.

### 8.3.2 Heteroscedasticity across clusters

One noteworthy thing about Figure 8.12 is that some speakers clearly have more inter-trial variability than others (compare, for example, speakers 15 and 11). This presence of inter-cluster differences in residual variability is called HETEROSCEDASTICITY. (The lack of heteroscedasticity—when residual intra-cluster variability is the same for all clusters—is called HOMOSCEDASTICITY.) Although the differences are not particularly severe in this case, we can still investigate whether they affect our inferences by incorporating them into our model. Conceptually speaking, this is a minor change to the structure of the model as depicted in Figure 8.9: the residual variance  $\sigma_y$  moves from being a shared  $\theta$  parameter to being a cluster-specific  $\mathbf{b}$  parameter. We present a Bayesian analysis (once again because methods for point-estimation are not readily available), with the following model:

$$\begin{aligned} \alpha, \beta_{\{1,2,3\}} &\sim \mathcal{N}(0, 10^5) \\ \mu_i &= \alpha + \beta_1 SO + \beta_2 MN + \beta_3 MS + b_i \\ b_i &\sim \mathcal{N}(0, \sigma_b) \\ y_{ij} &\sim \mathcal{N}(\mu_i, \sigma_{\mathbf{y},i}) \\ \log \sigma_b &\sim \mathcal{U}(-100, 100) \\ \log \sigma_{\mathbf{y},i} &\sim \mathcal{U}(-100, 100) \end{aligned}$$

Approximate 95% HPD confidence intervals and posterior modes obtained from sampling look as follows:

|          | lower bound | upper bound | posterior mode |
|----------|-------------|-------------|----------------|
| $\alpha$ | -1.79       | 0.21        | -0.64          |
| $SO$     | 4.96        | 9.29        | 7.69           |
| $MN$     | -2.5        | 2.04        | 0.1            |
| $MS$     | -1.56       | 4.92        | 3.01           |

Comparing these results with the point-estimate results obtained in the previous section, we see that failing to account for heteroscedasticity doesn't qualitatively change the conclusions of the model: there is still a strong association of clan origin with tone 1 slope, and there are no reliable effects of migration. Once again, similar inferences from multiple model specifications should strengthen your confidence in the conclusions obtained.

### 8.3.3 Multiple clusters per observation

One of the most exciting new developments in hierarchical modeling has been improvement in the computational treatment of cases where there are multiple classes of cluster to which each observation belongs. Consider the typical psycholinguistics experiment in speech perception or language comprehension, where each observation is derived from a particular participant reading or listening to a particular experimental stimulus which appears in a certain form. For example, the self-paced reading experiment of ?, described in Section 6.6.6 involved 58 subjects each reading 20 sentences (ITEMS), where each sentence could appear in any of four possible forms. The sample sentence is repeated below:

- (1) John {detests/babysits} the children of the musician who {is/are} generally arrogant and rude.

where there are two experimentally manipulated predictors: the type of verb used (implicit causality (IC) or non-IC), and the level of relative-clause attachment (high or low). This corresponds to a more complex hierarchical model structure, shown in Figure 8.13. In this figure, there are two cluster identity nodes  $i$  and  $j$ ; the subject-specific effects for the  $i$ -th subject are denoted by  $\mathbf{b}_{S,i}$ , and the item-specific effects for the  $j$ -th item are denoted by  $\mathbf{b}_{I,j}$ . This type of model is conceptually no different than the simpler hierarchical models we have dealt with so far. We illustrate by replicating the analysis of variance performed in Section 6.6.6 using a hierarchical linear model. Because the interaction between RC attachment level and verb type is of major interest to us, it is critical to us that we use an appropriate contrast coding (Section ??), using predictors  $V$  to represent verb type, with values 0.5 and -0.5 for IC and non-IC verb types, and  $A$  to represent attachment level, with values 0.5 and -0.5 for high and low attachment. We will allow different subject- and item-specific effects for each of the four possible conditions  $C$ . The hierarchical model can be compactly written as follows:

$$\begin{aligned}
 \mathbf{b}_{S,i} &\sim \mathcal{N}(\mathbf{0}, \Sigma_{\mathbf{b}_S}) \\
 \mathbf{b}_{I,j} &\sim \mathcal{N}(\mathbf{0}, \Sigma_{\mathbf{b}_I}) \\
 \mu_{ij} &= \alpha + \beta_V V + \beta_A A + \beta_{VA} VA + \mathbf{b}_{S,i}C + \mathbf{b}_{I,j}C \\
 y_{ijk} &\sim \mathcal{N}(\mu_{ij}, \sigma_{\mathbf{y}})
 \end{aligned}
 \tag{8.10}$$

We'll start by using point estimation of the model parameters with maximum likelihood. There is a considerable amount of detail in the resulting model so we present the complete printed representation of the fitted model from R:

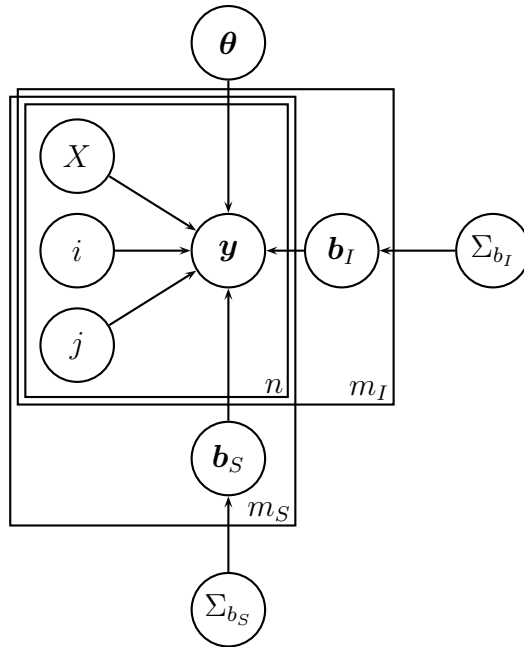


Figure 8.13: A conditional hierarchical model for probability distributions of the form  $P(Y|X, i)$ , with observations cross-classified into two classes of clusters

Linear mixed model fit by maximum likelihood

Formula:  $rt \sim V * A + ((C - 1) | subj) + ((C - 1) | item)$

Data: d

| AIC   | BIC   | logLik | deviance | REMLdev |
|-------|-------|--------|----------|---------|
| 12527 | 12648 | -6239  | 12477    | 12446   |

Random effects:

| Groups | Name        | Variance | Std.Dev. | Corr              |
|--------|-------------|----------|----------|-------------------|
| subj   | CIC high    | 11969.73 | 109.406  |                   |
|        | CIC low     | 18983.16 | 137.779  | 0.909             |
|        | CnonIC high | 23270.72 | 152.547  | 0.883 0.998       |
|        | CnonIC low  | 16472.44 | 128.345  | 0.988 0.963 0.946 |
| item   | CIC high    | 659.43   | 25.679   |                   |
|        | CIC low     | 563.45   | 23.737   | 1.000             |
|        | CnonIC high | 6062.47  | 77.862   | 0.271 0.271       |
|        | CnonIC low  | 3872.33  | 62.228   | 0.991 0.991 0.398 |

Residual 38503.03 196.222

Number of obs: 919, groups: subj, 55; item, 20

Fixed effects:

|             | Estimate | Std. Error | t value |
|-------------|----------|------------|---------|
| (Intercept) | 470.483  | 20.603     | 22.836  |
| V           | -33.708  | 16.777     | -2.009  |

|     |         |        |        |
|-----|---------|--------|--------|
| A   | -0.417  | 15.689 | -0.027 |
| V:A | -85.305 | 35.000 | -2.437 |

The negative estimates for  $\beta_V$  and  $\beta_A$  indicate that the IC-verb and high-attachment conditions respectively are associated with faster reading time at this region, though the effect of attachment level is very small.  $VA$  is positive in the IC/high and non-IC/low conditions, so the negative estimate for  $\beta_{VA}$  indicates that reading times at this region are faster in these two conditions.

We now turn to assessment of statistical significance. Once again resorting to our rule of thumb that with many more observations than estimated parameters (13 versus 919 in our case), the  $t$  statistic is distributed approximately as the standard normal, we see that the hierarchical analysis leads us to the same conclusions as the ANOVA of Section 6.6.6: there is a reliable interaction between verb type and RC attachment level in reading times at the first spillover region (*generally* in Example I). Importantly, this result is now obtained from a single hypothesis test rather than from separate by-subjects and by-items tests as has been the tradition in psycholinguistics for the past three decades.<sup>5</sup>

### Interpreting model parameter estimates

The ability to do a single hypothesis test when observations belong to multiple cross-cutting clusters is an advantage of the hierarchical analysis approach. However, where the hierarchical approach really shines is in obtaining a single model that can be inspected, interpreted, and used. From the subject-level variance parameters, we can see that there is much more intersubjective variation in reading speed for each condition (the subject-intercept standard deviations range from 111 to 154) than there is inter-item variability (standard deviations between 26 and 81). The residual trial-level variability is larger than both the subject- and item-level variability put together. Beyond this, however, there is something else worth noticing. The correlations between subject-level parameters for each condition are extremely high. That is, when a subject reads slowly in one condition, he or she reads slowly in *all* conditions. The correlations between item-level parameters are also high, except that there is much lower correlation between the implicit-causality, high-attachment condition and the rest of the conditions. This result could be illuminating if item-level parameter estimates were extracted and compared with the experimental materials themselves.

Turning to the shared parameters (“fixed effects”), we see that there is an overall 34-millisecond speed advantage for the implicit-causality verbs at this point, but no real overall effect of attachment level. Above and beyond these main effects, there is a large (85-ms) advantage for high-attaching RCs after IC verbs over low-attaching RCs after non-IC verbs. These estimates comport fairly well with those derived from the per-condition means obtained in Section 6.6.6, but the estimates obtained here put inter-subject and inter-item variation on equal footing, and are obtained automatically as part of the model-fitting process.

---

<sup>5</sup>Single hypothesis tests can also be done with separate by-subjects and by-items ANOVAs using the *min-F* test introduced by Clark (1973), this test is quite conservative and in practice is very rarely used.

## Fully Bayesian analysis

We can try a similar analysis using fully Bayesian techniques rather than the point estimate. We'll present a slightly simpler model in which the speaker- and item-specific variations do not depend on condition; this simpler type of model is often called a model with *random subject- and item-specific intercepts* in the literature. A Bayesian version of the more complex model of the previous section is left to the reader (see Exercise 8.7). Here is the model specification:

$$\begin{aligned}\alpha, \beta_{\{1,2,3\}} &\sim \mathcal{N}(0, 10^5) \\ \log \sigma_{b_S} &\sim \mathcal{U}(-100, 100) \\ \log \sigma_{b_I} &\sim \mathcal{U}(-100, 100) \\ \log \sigma_{b_y} &\sim \mathcal{U}(-100, 100) \\ b_{S,i} &\sim \mathcal{N}(0, \sigma_{b_S}) \\ b_{I,j} &\sim \mathcal{N}(0, \sigma_{b_I}) \\ \mu_{ij} &= \alpha + \beta_1 V + \beta_2 A + \beta_3 VA + b_{S,i} + b_{I,j} \\ y_{ijk} &\sim \mathcal{N}(\mu_{ij}, \sigma_y)\end{aligned}$$

Note the close resemblance to the previous model specification in Equation 8.10; the two differences are the addition of the top four lines representing priors over the shared parameters  $\alpha$  and  $\beta_{\{1,2,3\}}$ , and the simplified  $b_{S,i}$  and  $b_{I,j}$  since we only have subject- and item-specific intercepts now.

Sampling from the posterior gives us the following 95% HPD confidence intervals and posterior modes for the effects of  $V$ ,  $A$ , and the interaction  $VA$ :

|      | lower bound | upper bound | posterior mode |
|------|-------------|-------------|----------------|
| $V$  | -54.09      | -4.19       | -37.11         |
| $A$  | -28.33      | 20.65       | -9.47          |
| $VA$ | -136.26     | -41.52      | -50.77         |

Once again, there is broad agreement between the point estimates obtained earlier in this section and the Bayesian HPD confidence intervals. Most notably, there is (a) strong evidence of an overall trend toward faster reading in this region for the IC verbs; and (b) even stronger evidence for an interaction between IC verb type and attachment level. One should believe an apparent trend in the data more strongly if the same trend is confirmed across multiple statistical analyses, as is the case here.

## 8.4 Hierarchical generalized linear models

We now shift from linear models to the broader case of generalized linear models, focusing on logit models since they (along with linear models) are the most widely used GLM in the

study of language. We move from generalized linear models (GLMs) to hierarchical GLMs by adding a stochastic component to the linear predictor (c.f. Equation 6.1):

$$\eta = \alpha + (\beta_1 + b_{i1})X_1 + \cdots + (\beta_n + b_{in})X_n \quad (8.11)$$

and assume that the cluster-specific parameters  $\mathbf{b}$  themselves follow some distribution parameterized by  $\Sigma_{\mathbf{b}}$ .

We then follow the rest of the strategy laid out in Section 6.1 for constructing a GLM: choosing a link function  $\eta = l(\mu)$ , and then choosing a function for noise around  $\mu$ .

### 8.4.1 Hierarchical logit models

In a hierarchical logit model, we simply embed the stochastic linear predictor in the binomial error function (recall that in this case, the predicted mean  $\mu$  corresponds to the binomial parameter  $\pi$ ):

$$P(y; \mu) = \binom{n}{yn} \mu^{yn} (1 - \mu)^{(1-y)n} \quad (\text{Binomial error distribution}) \quad (8.12)$$

$$\log \frac{\mu}{1 - \mu} = \eta \quad (\text{Logit link}) \quad (8.13)$$

$$\mu = \frac{e^\eta}{1 + e^\eta} \quad (\text{Inverse logit function}) \quad (8.14)$$

### 8.4.2 Fitting and interpreting hierarchical logit models

As with hierarchical linear models, the likelihood function for a multi-level logit model must marginalize over the cluster-level parameters  $\mathbf{b}$  (Equation 8.5). We can take either the maximum-likelihood approach or a Bayesian approach. Unlike the case with hierarchical linear models, however, the likelihood of the data  $P(\mathbf{y}|\theta, \Sigma_{\mathbf{b}})$  (marginalizing over cluster-level parameters  $\mathbf{b}$ ) cannot be evaluated exactly and thus the MLE must be approximated (Bates, 2007, Section 9). The tool of choice for approximate maximum-likelihood estimation is once again the `lme4` package in R.<sup>6</sup>

### 8.4.3 An example

We return to the dataset of Bresnan et al. (2007), illustrated by the alternation

- (2) Susan gave *toys to the children*. (PP realization of recipient)
- (3) Susan gave *the children toys*. (NP realization of recipient)

---

<sup>6</sup>The recommended approximations to maximum likelihood are Laplacian approximation (see, e.g., Robert and Casella (2004, Section 3.4)) and adaptive Gaussian quadrature; the former is available in `lme4` and the recommended default, but the latter isn't (yet).

To illustrate the approach, we construct a model with the length, animacy, discourse accessibility, pronominality, and definiteness of both the recipient and theme arguments as predictors, and with verb as a random effect. We use log-transformed length predictors (see Section 6.7.4 for discussion).

### Defining the model

We arbitrarily denote length, animacy, discourse status, pronominality, and definiteness of the theme with the variables  $L_T, A_T, S_T, P_T, D_T$  respectively, and those properties of the recipient as  $L_R, A_R, S_R, P_R, D_R$ .<sup>7</sup> We can now write our hierarchical model as follows:

$$\begin{aligned}
 b_i &\sim N(0, \sigma_b) \\
 \eta_i &= \alpha + \beta_{L_T} L_T + \beta_{A_T} A_T + \beta_{S_T} S_T + \beta_{P_T} P_T + \beta_{D_T} D_T \\
 &\quad + \beta_{L_R} L_R + \beta_{A_R} A_R + \beta_{S_R} S_R + \beta_{P_R} P_R + \beta_{D_R} D_R + b_i \\
 \pi_i &= \frac{e_i^\eta}{1 + e_i^\eta} \\
 y_{ij} &\sim \text{Binom}(1, \pi_i)
 \end{aligned} \tag{8.15}$$

(Note that we could equally specify the last line as a Bernoulli distribution:  $y_{ij} \sim \text{Bern}(\pi)$ .) We arbitrarily consider prepositional-object (PO) realization of the recipient as the “successful” outcome (with which positive contributions to the linear predictor  $\eta$  will be associated). Approximate maximum-likelihood estimation gives us the following parameter estimates: The fixed-effect coefficients, standard errors, and Wald  $z$ -values can be interpreted as normal in a logistic regression (Section 6.7.1). It is important to note that there is considerable variance in verb-specific preferences for PO versus DO realizations. The scale of the random effect is that of the linear predictor, and if we consult the logistic curve we can see that a standard deviation of 2.33 means that it would be quite typical for the magnitude of this random effect to be the difference between a PO response probability of 0.1 and 0.5.

We now turn our attention to the shared model parameters. The following properties are associated with PO outcomes:

- Longer recipients
- Inanimate recipients
- Discourse-new recipients
- Non-pronominal recipients
- Indefinite recipients
- Shorter themes
- Animate themes
- Discourse-old themes
- Pronominal themes
- Definite themes

---

<sup>7</sup>To simplify model interpretation, I have reduced the tripartite distinction of Bresnan et al. of discourse status as **given**, **accessible**, and **new** into a binary distinction of **given** versus **new**, with **accessible** being lumped together with **new**.



```

Generalized linear mixed model fit by the Laplace approximation
Formula: RealizationOfRecipient ~ log(LengthOfRecipient) + log(LengthOfTheme) + Ani
Data: dative
AIC   BIC logLik deviance
1473 1546 -724.3    1449
Random effects:
Groups Name          Variance Std.Dev.
Verb  (Intercept) 5.4344  2.3312
Number of obs: 3263, groups: Verb, 75

Fixed effects:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      2.3198    0.6616   3.506 0.000454 ***
log(LengthOfRecipient)  1.3145    0.1522   8.636 < 2e-16 ***
log(LengthOfTheme)   -1.1703    0.1067  -10.967 < 2e-16 ***
AnimacyOfRecinanimate  2.1357    0.2533   8.430 < 2e-16 ***
AnimacyOfThemeinanimate -0.9231    0.4980  -1.854 0.063811 .
DiscStatusOfRecnew    1.3263    0.2060   6.437 1.22e-10 ***
DiscStatusOfThemew    -1.7650    0.2722  -6.485 8.85e-11 ***
PronomOfRecpronominal -1.5449    0.2304  -6.706 2.00e-11 ***
PronomOfThemepronominal  2.2025    0.2630   8.375 < 2e-16 ***
DefinOfRecindefinite  0.7963    0.2008   3.965 7.33e-05 ***
DefinOfThemeindefinite -1.0898    0.1984  -5.493 3.96e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Figure 8.14: Point estimate for a hierarchical logit model of the dative alternation

There is a clear trend here: those properties of the theme that favor PO outcomes are the reverse of those properties that favor DO outcomes. This raises the linguistically interesting possibility that there is a unified set of principles that applies to word ordering preferences in the English postverbal domain and which is sensitive to high-level properties of constituents such as length, discourse status, and so forth, but *not* to specific combinations of these properties with the grammatical functions of the constituents. This possibility is followed up on in Exercise 8.9.

### Inferences on cluster-specific parameters

Because of this considerable variance of the effect of verb, it is worth considering the inferences that we can make regarding verb-specific contributions to the linear predictor. One way of doing this would be to look at the conditional modes of the distribution on verb-specific effects  $\mathbf{b}$ , that is to say the BLUPs (Section 8.2.1). There is a disadvantage to this

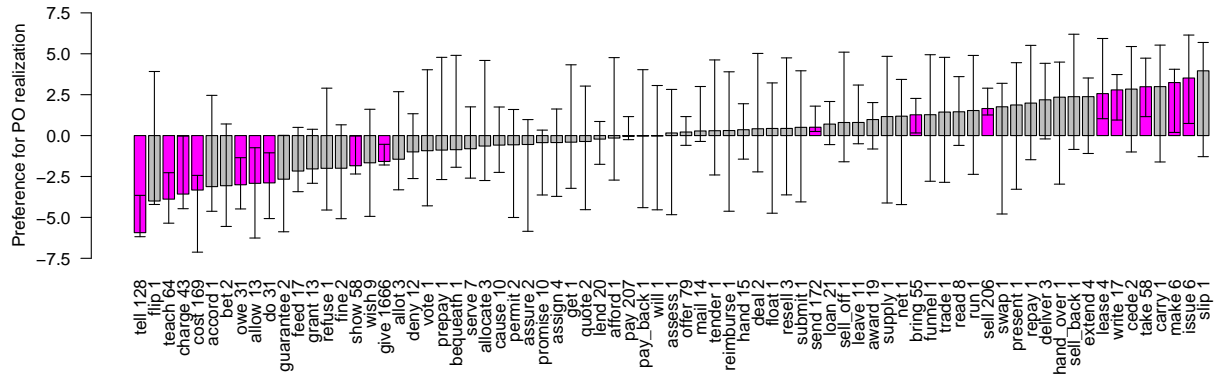


Figure 8.15: Verb-specific preferences in analysis of the **dative** dataset. 95% HPD confidence intervals are plotted on each preference. Verbs for which the 95% HPD interval is entirely on one side of the origin have their bars plotted in magenta.

approach, however: there is no easy way to assess our degree of confidence in the conditional modes. Another option is to use a fully Bayesian approach and plot posterior modes (of  $P(\mathbf{b}|y, \Sigma_{\sigma_b}, \Sigma_{\theta})$ , which is different from the BLUPs) along with confidence intervals. This is the approach taken in Figure 8.15, using uniform priors on all the shared parameters as well as on  $\log \sigma_b$ . On the labels axis, each verb is followed by its SUPPORT: the number of instances in which it appears in the **dative** dataset. For most verbs, we do not have enough information to tell whether it truly has a preference (within the model specification) toward one realization or the other. However, we do have reliable inferences some verbs: for the most part, those with large support and/or with posterior modes far from 0.<sup>8</sup> We can see that *tell*, *teach*, *charge*, and *show* are strongly biased toward the double-object construction, whereas *loan*, *bring*, *sell*, and *take* are biased toward the prepositional-object construction.

These results are theoretically interesting because the dative alternation has been at the crux of a multifaceted debate that includes:

- whether the alternation is meaning-invariant;
- if it is not meaning-invariant, whether the alternants are best handled via constructional or lexicalist models;
- whether verb-specific preferences observable in terms of raw frequency truly have their locus at the verb, or can be explained away by other properties of the individual clauses at issue.

Because verb-specific preferences in this model play such a strong role despite the fact that many other factors are controlled for, we are on better footing to reject the alternative raised

<sup>8</sup> This is not the whole story, however: comparing *deny* with *promise*, the former has both larger support and a more extreme posterior mode, but it is the latter that has an HPD confidence interval that is closer to not including 0.

by the third bullet above that verb-specific preferences can be entirely explained away by other properties of the individual clauses. Of course, it is always possible that there are other explanatory factors correlated with verb identity that will completely explain away verb-specific preferences; but this is the nature of any type of scientific explanation. (This is also a situation where controlled, designed experiments can play an important role by eliminating the correlations between predictors.)

#### 8.4.4 Model comparison & hypothesis testing

The framework for hypothesis testing in hierarchical generalized linear models is similar overall to that for hierarchical linear models as described in Section 8.3.1. For model comparisons involving the same shared-parameter structure but nested cluster-specific parameter structures, likelihood-ratio tests are conservative. For the assessment of the significance of a single shared parameter estimate  $\hat{\beta}_i$  against the null hypothesis that  $\beta_i = 0$ , the Wald  $z$ -statistic (Section 6.8.1), which is approximately standard-normal distributed under the null hypothesis, can be used. In Figure 8.14, for example, the  $z$ -statistic for log of recipient length is  $\frac{1.31}{0.15} = 8.64$ , which is extremely unlikely under the null hypothesis.

To simultaneously assess the significance of the contribution of more than one shared parameter to a model, a likelihood-ratio test is most appropriate. Once again, however, in hierarchical models this test may be anti-conservative, as described in Section 8.3.1, so caution should be used in interpreting apparently significant results. As an example of how this test can be useful, however, let us consider an alternative model of the dative alternation in which the tripartite discourse status of recipients and themes into given, accessible, and new (Collins, 1995). This difference introduces two new parameters into the model. Twice the difference in the log-likelihoods of the original and the updated model is 3.93; the cumulative distribution function for  $\chi_2^2$  at this point is 0.86, giving us a best-case  $p$ -value of 0.14, so we can safely state that we don't have sufficient evidence to adopt the tripartite distinction over the bipartite given/new distinction used earlier.

#### 8.4.5 Assessing the overall performance of a hierarchical logit model

We conclude the chapter with a brief discussion of assessing overall performance of hierarchical logit models. As with any probabilistic model, the data likelihood is an essential measure of model quality, and different candidate models can be compared by assessing the likelihood they assign to a dataset. Cross-validated or held-out likelihood can be used to alleviate concerns about overfitting (Section 2.11.5). It is also useful to visualize the model's performance by plotting predictions against empirical data. When assessing the fit of a model whose response is continuous, a plot of the residuals is always useful. This is not a sensible strategy for assessing the fit of a model whose response is categorical. Something that is often done instead is to plot *predicted probability* against *observed proportion* for some binning of the data. This is shown in Figure 8.16 for 20 evenly-spaced bins on the  $x$ -axis, with the point representing each bin of size proportionate to the number of observations summarized in that point. There is a substantial amount of information in this plot: the model is quite

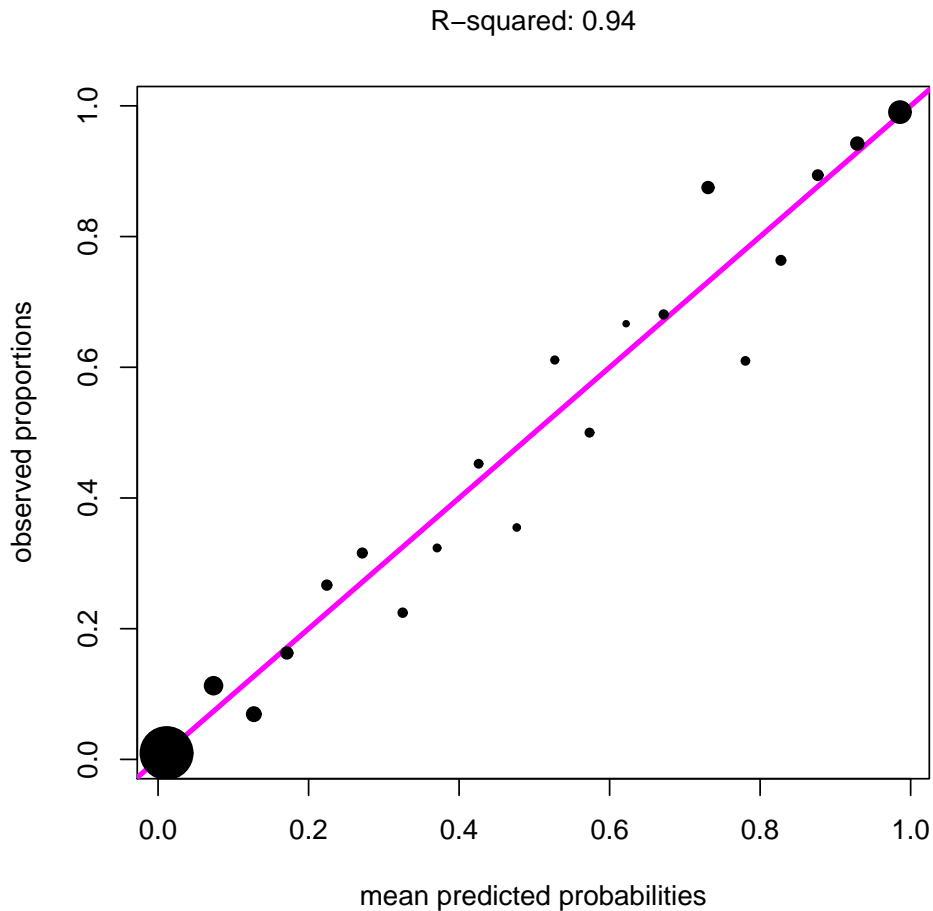


Figure 8.16: The fit between predicted probabilities and observed proportions for each 5% of predicted probability for our model

certain about the expected outcome for most observations, and the worst-outlying bins are between predicted probability of 0.7 and 0.8, but contain relatively little data. Producing this visualization could be followed up by examining those examples to see if they contain any important patterns not captured in the model. Probability-proportion plots can also be constructed using cross-validation (Exercise 8.12).

## 8.5 Further Reading

Hierarchical models are an area of tremendous activity in statistics and artificial intelligence. There is good theoretical coverage (and some examples) of hierarchical generalized linear models in Agresti (2002, Chapter 12). Pinheiro and Bates (2000) is an important book on theory and practice for linear and non-linear hierarchical models from the frequentist per-

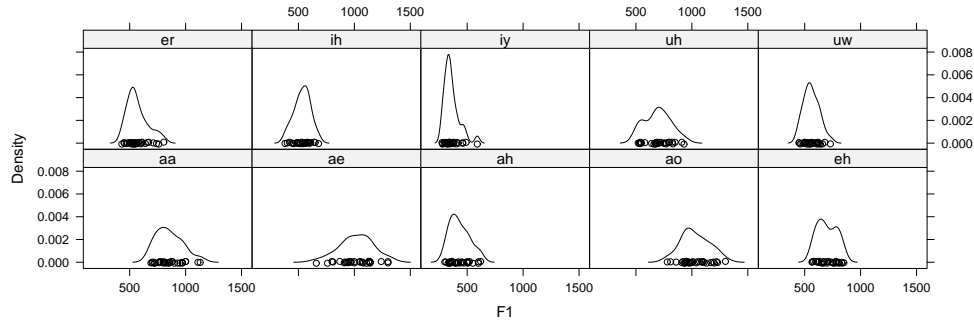


Figure 8.17: Density plots of children’s F1 formant recordings from Peterson & Barney

spective. There is also a bit of R-specific coverage in Venables and Ripley (2002, Section 10.4) which is useful to read as a set of applied examples, but the code they present uses penalized quasi-likelihood estimation and this is outdated by `lme4`. A more recent and comprehensive text for hierarchical regression models is Gelman and Hill (2007), which focuses the Bayesian perspective but is practically oriented, and includes coverage of both `lme4` and `BUGS`. At the time of writing, this is probably the single best place to turn to when learning the practicalities of working with hierarchical models for the analysis of complex datasets.

## 8.6 Exercises

### Exercise 8.1: Defining a hierarchical model

The F1 formant levels of children’s vowels from Peterson & Barney’s dataset tend not to be normally distributed, but are often right-skewed (Figure 8.17). Define a hierarchical probabilistic model of F1 formant values for the vowel [a] in which the speaker-specific variation component  $b_i$  are log-normally distributed—that is, if  $\log b_i = x_i$ , then the  $x_i$  are normally distributed with some mean and standard deviation. Write the hierarchical model in the style of Equation 8.2. Choose parameters for your model by hand that generate data that looks qualitatively like that of Figure 8.17.

### Exercise 8.2: Shrinkage

Consider the simple point estimate hierarchical model of [a] first-formant (F1) data from Peterson and Barney (1952) obtained in Section 8.2.1. In this model, calculate the joint log-likelihood of  $\hat{\mathbf{b}}$  and  $\mathbf{y}$ —that is, of speaker-specific averages and trial-specific observations—using as  $\hat{\mathbf{b}}$  (i) the average recorded F1 frequency for each speaker, and (ii) the conditional modes, or BLUPs, obtained from the hierarchical model. Is the joint log-likelihood higher in (i) or (ii)?

### Exercise 8.3

Replicate the posterior inference procedure in Section 8.2.2 for F2 formant frequencies

from Peterson and Barney (1952).

#### **Exercise 8.4: Priors on $\sigma_y$**

In Section 8.2.2, we used a prior that was locally uniform on the log of the variance parameter  $\sigma_y$ . Another alternative would be to use a prior that was locally uniform on  $\sigma_y$  itself. Check to see how much of a difference, if any, this alternative would have on posterior inference over  $\mu$  and  $\Sigma_b$ . (You probably will want to do several simulations for each type of model to get a sense of how much variation there is over samples in each case.)

#### **Exercise 8.5: Hierarchical Bayesian linear models**

Use R and BUGS together to replicate the analysis of Section 8.3.1, but with homoscedastic intra-cluster variation. Do the Bayesian confidence intervals on effects of northward migration, southward migration, and clan origin look similar to those obtained using point estimation?

#### **Exercise 8.6: Hypothesis testing for random effects**

Using the method of maximum likelihood via `lmer()`, together with the likelihood-ratio test, conduct a hypothesis test for the implicit-causality data of Section 8.3.3 on whether subject- and item-specific effects of experimental condition (as opposed to just subject- and item-specific intercepts) significantly improve the likelihood of a hierarchical model. That is, build a model where the only cluster-specific parameters are by-subject and by-item intercepts, and compare its log-likelihood to the full model in which all experimental conditions interact. What are your conclusions?

#### **Exercise 8.7: Hierarchical Bayesian linear models with crossed effects**

Replicate the fully Bayesian analysis presented in Section 8.3.3, but with condition-specific inter-subject and inter-item variation. Write down the complete model specification, implement it with JAGS, and compare the resulting inferences about model parameters with those from the point-estimate analysis and from the simpler Bayesian model. Be sure to consider both inferences about shared parameters and about the parameters governing cross-cluster variation!

#### **Exercise 8.8: Transforms of quantitative predictors in mixed-effect models**

Re-run the `datave.glm` regression from Section 8.4.3, but use raw constituent lengths rather than log-transformed lengths. Compute cross-validated likelihoods (Section 2.11.5) to compare the performance of this model and of the original model? Which approach yields higher log-likelihood?

#### **Exercise 8.9: Simplifying a model based on linguistic principles and model comparison**

Define a simpler version of the model in Section 8.4.3 in which the effects of each of constituent length, animacy, discourse status, pronominality, and definiteness must be equal and opposite for recipients versus themes. Fit the model using approximate maximum-likelihood estimation, and compare it to the original model using the likelihood ratio test. Can you

safely conclude that the effects of these factors truly are equal and opposite? (**Hint:** the easiest way to construct the simpler model is to define new quantitative predictors that express the summed influence of the property from both recipient and theme; see Exercise 6.10. Also keep in mind that the likelihood-ratio test can be anti-conservative for shared (“fixed-effect”) parameters in hierarchical models.)

### Exercise 8.10: Mixed-effects logit models and magnitudes of parameter estimates

Unlike with mixed-effects linear models, accounting for a cluster-level variable in a mixed-effects logit model can systematically change the magnitude of the parameter estimates for fixed effects. (**Warning:** this may be a pretty hard problem.)

1. Re-run the `dative.glm` regression from Section 8.4.3 as a standard logistic regression model, completely omitting the random effect of verb, but replacing it with a fixed effect of the verb’s SEMANTIC CLASS). Do the magnitudes of most of the fixed-effect coefficients (intercept excluded) increase or decrease? Can you think of any reason why this would happen?
2. Test your intuitions by constructing a simple population with two clusters (call the factor `Cluster` with levels `C1,C2`) and a single two-level fixed effect (call the factor `Treatment` with levels `A,B`). Assume the underlying model involves the following linear predictor:

$$\eta = -1 + 2X + bZ$$

where  $X$  is a dummy indicator variable that is active when treatment level is `B`,  $Z$  is the cluster variable, and  $b$  is 1 for cluster 1 and -1 for cluster 2. Generate a dataset consisting of the following cell counts:

|   | C1  | C2  |
|---|-----|-----|
| A | 250 | 250 |
| B | 250 | 250 |

using a logit model, and fit (1) a standard logistic regression with only `Treatment` as a fixed effect, and (2) a mixed-effects logistic regression. How does the estimation of the fixed effect change between models (1) and (2)?

3. Repeat your controlled experiment from (b) above, except this time use linear models (classic and mixed-effects) where the noise has standard deviation 1. Does the same change in the estimated effect of `Treatment` occur as in the logit model?

### Exercise 8.11: Different priors on verb-specific effects

1. Use kernel density estimation to estimate the posterior density of the BLUPs for verbs in the dative alternation from the hierarchical logit model of Section 8.4.3. Plot this

estimated density, and overlay on it a normal density with mean 0 and standard deviation  $\hat{\sigma}_{\mathbf{b}}$ . Do the BLUPs look like they follow this normal distribution? Choose another distribution for  $\mathbf{b}$  in this model and implement it in BUGS. (Be forewarned that this is conceptually relatively simple but computationally rather intensive; loading the model and sampling from the posterior are likely to take several minutes or hours even on a new processor.) Plot the posterior inferences on  $\mathbf{b}$  as in Figure 8.15 and compare them to the results of the original model. Are the results considerably different?

**Exercise 8.12: Cross-validated likelihood and probability-proportion plot**

Reconstruct Figure 8.16 using ten-fold cross-validation to obtain predicted probabilities. Does the resulting fit look better? Worse? Also compare the cross-validated likelihood to the original model's likelihood. Does the model seem substantially overfit?



## Chapter 9

# Dimensionality Reduction and Latent Variable Models

### 9.1 Gaussian mixture models

As a case study to motivate our first latent-variables model, we consider the problem of how infants learn the phonological categories of their native language. There is ample evidence that phonological-category learning involves a combination of both innate bias and experience. There are some phonetic distinctions which cannot be reliably identified by adult native speakers of languages that lack them (e.g., alveolar [d] versus retroflex [ɖ] for English speakers, or [r] versus [ɺ] for Japanese speakers; Werker and Tees, 1984; Kuhl et al., 2006, *inter alia*), suggesting the power of innate bias. Likewise, it has more recently become clear that there are some phonetic distinctions which cannot be reliably identified by younger infants but which can be reliably identified not only by adult native speakers of languages which possess the distinction but also by older infants being raised in those language environments (e.g., syllable-initial [n] versus [ŋ] in Filipino language environments; citealp narayan-et al: Explicit models of phonological category learning could potentially be of considerable value in gaining insight into how such learning takes place across languages.

An appreciation for some of the challenges inherent in the phonological category learning problem can be appreciated by considering inter-speaker variation in the realization of different categories. Figure 9.1 presents sample data drawn from multivariate Gaussian approximations of the realization of the vowels [i], [ɪ], [e], [ɛ] (as in *beet*, *bit*, *bait*, and *bet* respectively), for two native Canadian English speakers, plotted by their F1 and F2 formants as well as vowel durations respectively. To minimize gross inter-speaker variation, each dimension has been standardized according to the speaker's overall distributions in these three dimensions. Two difficulties stand out in this graph. First, whereas the first speaker has clear separation among the four vowels, the separation for the second speaker is much less clear; it is really essential to take F1 into account even to separate [ɛ] from the other three vowels, and the separation among the remaining three vowels is poor. Second, the distinctions among the vowels are not entirely robust across speakers: for example, whereas for the first speaker [e]

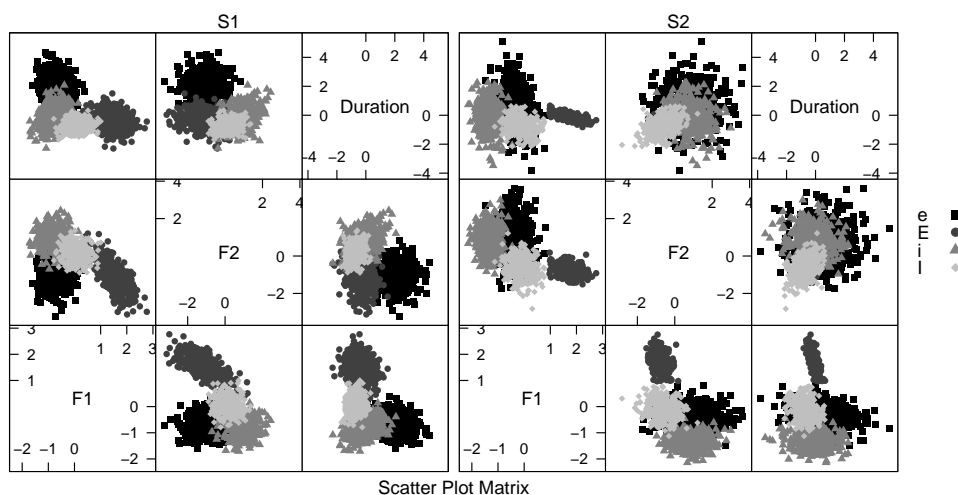


Figure 9.1: F1 frequency, F2 frequency, and duration of vowels [i],[ɪ],[e],[ɛ] for two speakers of English. Each of the three variables is standardized for each speaker.

tends to have lower F2 than [ɪ], for the second speaker it is [ɪ] that tends to have lower F2 than [e]. Figure 9.2 shows samples from all four vowels mixed across nineteen speakers of English.

We will formalize the learning problem faced by the infant as follows: given a set of observations in a multidimensional phonetic space, draw inferences about (i) the proper grouping of the observations into categories; and (ii) the underlying abstract representations of those categories (e.g., what new observations drawn from each category are likely to look like). Every possible grouping of observations  $\mathbf{y}$  into categories represents a PARTITION  $\Pi$  of the observations  $\mathbf{y}$ ; let us denote the parameters determining the categories' underlying representations as  $\boldsymbol{\theta}$ . In probabilistic terms, our learning problem is to infer the two following probability distributions:

$$P(\Pi|\mathbf{y}) \quad (\text{the distribution over partitions given the data; and}) \quad (9.1)$$

$$P(\boldsymbol{\theta}|\mathbf{y}) \quad (\text{the distribution over category properties given the data.}) \quad (9.2)$$

To complete the formalization of the learning problem, we need to specify the representations  $\boldsymbol{\theta}$  of the underlying categories, as well as a method for inferring the two probability distributions above. For the present problem, we will assume that each phonological category consists of a multivariate Gaussian distribution in the three-dimensional space of F1 frequency, F2 frequency, and vowel duration; hence the representations are the means  $\boldsymbol{\mu}$  and covariance matrices  $\boldsymbol{\Sigma}$  of multivariate normal distributions (see Section 3.5). This specification of category representations gives us the beginning of a generative model of our observations: each observation is drawn independently from the Gaussian distribution associated with one underlying category.

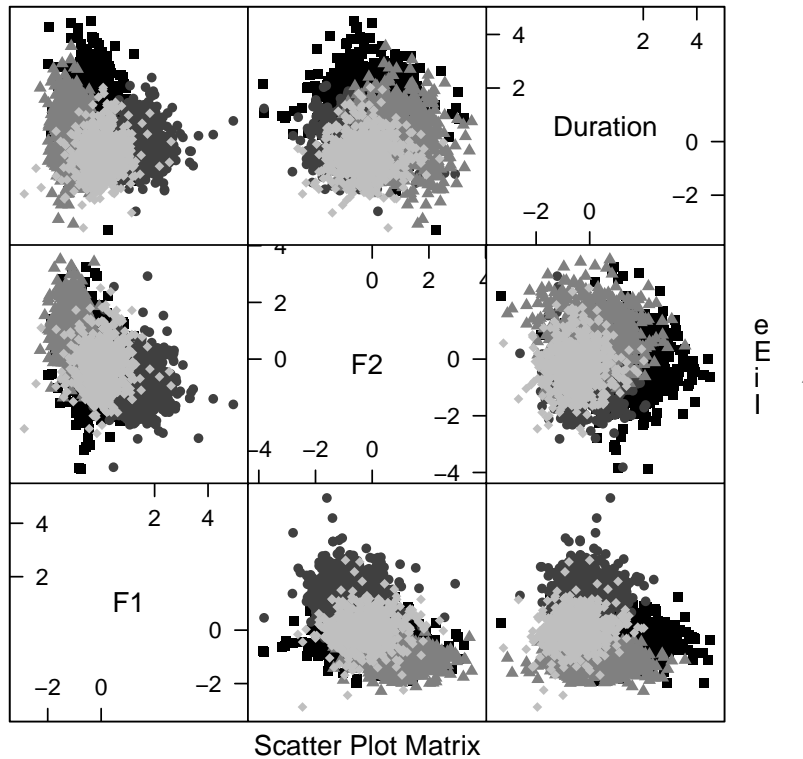


Figure 9.2: F1 frequency, F2 frequency, and duration of vowels [i],[ɪ],[e],[ɛ], mixed from 19 speakers of English.

Thus far, the model is formally identical to the models examined in Chapter 6, where the category  $i$  to which each observation belongs could be treated as a categorical predictor variable, and our model specification is of the conditional probability distribution  $P(Y|i)$ . The crucial difference, however, is that in the present setting, the category identity of each observation is unknown—it is a LATENT VARIABLE which needs to be inferred. This means that the current specification is incomplete: it does not give us a probability distribution over possible observations in phonetic space. To complete the generative model, we need to specify the multinomial probabilities  $\phi$  that an observation will be drawn from each underlying category. This gives us the following generative model:

$$i \sim \text{Multinom}(\phi) \tag{9.3}$$

$$y \sim \mathcal{N}(\mu_i, \Sigma_i) \tag{9.4}$$

$$\tag{9.5}$$

In the probabilistic setting, the complete underlying properties to be inferred are thus  $\theta = \langle \phi, \mu, \Sigma \rangle$ . This model is known as a MIXTURE OF GAUSSIANS, since the observations are drawn from some mixture of individually Gaussian distributions. An illustration

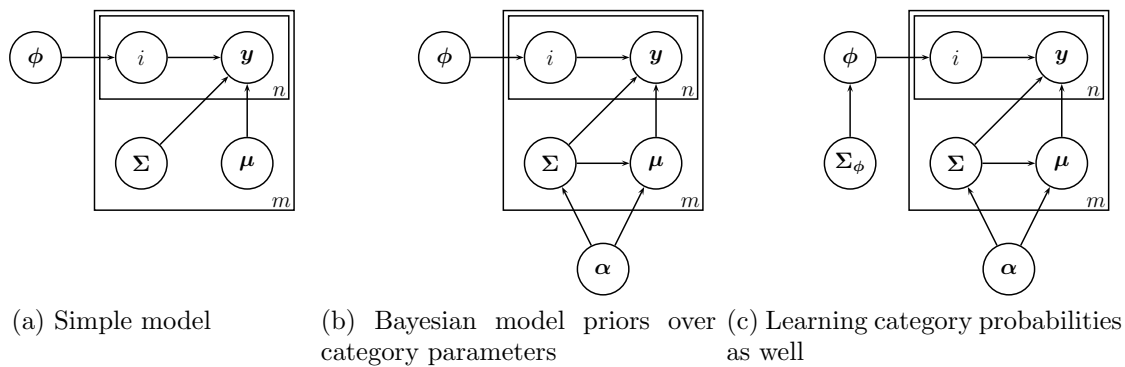


Figure 9.3: Graphical model for simple mixture of Gaussians

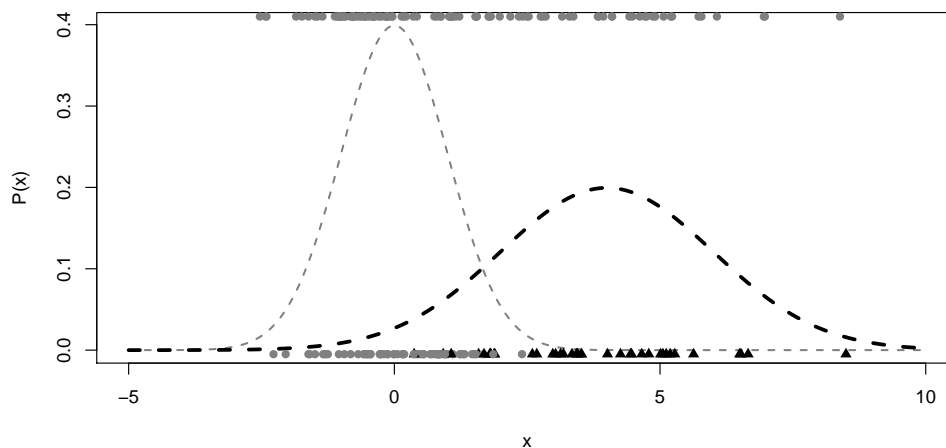


Figure 9.4: The generative mixture of Gaussians in one dimension. Model parameters are:  $\phi_1 = 0.35, \mu_1 = 0, \sigma_1 = 1, \mu_2 = 4, \sigma_2 = 2$ .

of this generative model in one dimension is given in Figure 9.4, with the Gaussian mixture components drawn above the observations. At the bottom of the graph is one sample from this Gaussian mixture in which the underlying categories are distinguished; at the top of the graph is another sample in which the underlying categories are not distinguished. The relatively simple supervised problem would be to infer the means and variances of the Gaussian mixture components from the category-distinguished sample at the bottom of the graph; the much harder unsupervised problem is to infer these means and variances—and potentially the number of mixture components!—from the category-undistinguished sample at the top of the graph. The graphical model corresponding to this unsupervised learning problem is given in Figure 9.3a.

### 9.1.1 Inference for Gaussian mixture models

Let us temporarily set aside the problem of inferring the number of underlying categories and treat it as solved (we will return to this problem in Chapter ??). Even so, we are faced with a challenging problem. If there are  $n$  observations and  $m$  categories, then there are  $m^n$  logically possible partitions, and it turns out that there is no way to get around this exponentially large number of partitions to find exact solutions for the two probabilities of interest, given in Equations (9.1) and (9.2). Furthermore, it isn't even possible to try obtaining a point estimate of Equation (9.2) through hill-climbing on a continuous surface, as we did with (for example) log-linear models in Chapter 6, because the partitions are discrete.

Additionally, there is a subtle but important property of the mixture-of-Gaussians problem that makes pure maximum-likelihood techniques deeply problematic. Consider a partition in which one category contains exactly one observation (in the case of Figure 9.4, for example, for the sample on the top of the graph let the rightmost point belong to one category, and let all other points belong to the other category). Now let the mean of that category be the observation itself. As the variance of that category approaches zero, the likelihood of that observation from that category approaches infinity (this is a case of the bias in the maximum-likelihood estimate of the variance, which we saw in Section 4.3.2). Therefore, any “maximum-likelihood” solution to this problem will involve categories which are centered around single observations and which have zero variance.

This, therefore, is a very clear case where we need to inject prior knowledge in order to learn anything meaningful about category inventories at all. At a minimum, it is essential that this prior knowledge enforces the constraint that category variances approaching zero are forbidden. This requirement makes the Bayesian framework especially appealing: the constraint on category variances can be formalized by letting the prior probability of each category's parameters go to zero as the category variance itself goes to zero. Within the Bayesian framework, we can use sampling techniques to approximate the posterior distributions of interest in Equations (9.1) and (9.2).

#### A first Gaussian mixture model

We begin by simplifying the problem in assuming that category probabilities  $\phi$  are known and equal:  $\phi_1 = \phi_2 = \phi_3 = \phi_4 = \frac{1}{4}$ . These category probabilities serve as a uniform prior distribution over partitions. However, we still need to put priors over the individual category means  $\mu$  and covariances  $\Sigma$  in order to draw inferences about them. As always, there are many choices here; for convenience, we choose priors that are widely used for Bayesian learning of normal distributions with unknown mean and variance. This prior is specified in two parts,  $P(\Sigma)$  and  $P(\mu|\Sigma)$ :

$$\begin{aligned}\Sigma_i &\sim \mathcal{IW}(\Sigma_0, \nu) \\ \mu_i|\Sigma &\sim \mathcal{N}(\mu_0, \Sigma_i/A)\end{aligned}$$

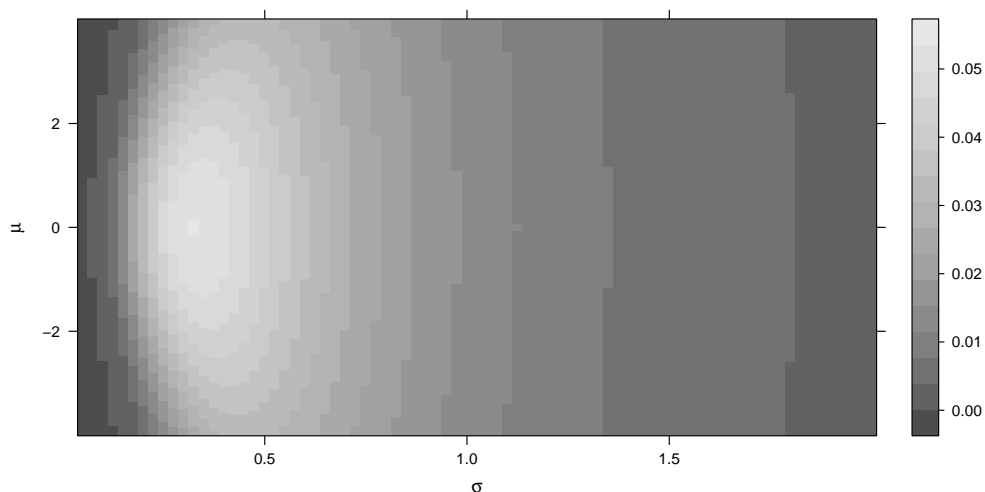


Figure 9.5: The joint distribution over component variances and means for a 1-dimensional Gaussian mixture model

where  $A$  is a scaling parameter and, as in Chapter 8,  $\mathcal{IW}$  signifies the inverse Wishart distribution (Chapter B). This formulation may seem complex at first glance, but it is actually fairly straightforward. The first line states that the prior distribution on the covariance matrix is centered around some matrix  $\Sigma_0$  which is treated as if it had been estimated from  $\nu$  samples. The second line states that the prior distribution on the mean is normal and centered around  $\mu_0$ , with covariance matrix proportional (by some factor  $1/A$ ) to the covariance matrix  $\Sigma_i$  for observations. Placing dependence of the prior for the mean on the prior for the covariance matrix may seem counter-intuitive, but there are two good reasons to do it. First, this prior is conjugate to the likelihood of observed data within the cluster.<sup>1</sup> Second, according to this prior, any indication that there is low variability across observations ( $\Sigma_i$  small) is also an indication that we can estimate the mean with relatively high certainty ( $\Sigma_i/A$  small). This is exactly the way that the variance and mean are related in inferences from observed data: when observed data are broadly distributed, we estimate the variance to be large and have low certainty about the mean, but observed data are tightly distributed, we estimate the variance to be small and have high certainty about the mean.

Figure 9.5 is a visualization of this conjugate joint prior distribution over  $\Sigma$  and  $\mu$  for univariate data (so that  $\Sigma$  can be represented as a scalar,  $\sigma$ ). Note that this distribution drops off rapidly to zero near  $\sigma = 0$ .

### Sampling from the posterior distribution

With this setup, it is straightforward to use Gibbs sampling (a type of Markov-chain Monte Carlo technique; see Section XXX) to draw samples from the joint posterior distribution

<sup>1\*\*\*</sup>Reference Gelman et al. for better treatment of this conjugacy.

$P(\Pi, \boldsymbol{\theta} | \mathbf{y})$ , and to use these samples as the basis for any inferences we wish to make. Gibbs sampling here can be taken as alternating between sampling of a partition given model parameters from  $P(\Pi | \boldsymbol{\theta}, \mathbf{y})$  and sampling of model parameters given a partition from  $P(\boldsymbol{\theta} | \Pi, \mathbf{y})$ . Because observations are conditionally independent of one another given model parameters (see Figure 9.3b), sampling a partition simply involves sampling a category  $i$  for each instance  $y_j$  conditioned on model parameters from  $P(i | y_j, \boldsymbol{\theta})$ , which can be rewritten using Bayes' rule as

$$P(i | y_j, \boldsymbol{\theta}) = \frac{P(y_j | i, \boldsymbol{\theta})}{\sum_i P(y_j | i, \boldsymbol{\theta})} \quad (9.6)$$

In our case, computing the numerator and each term of the sum in the denominator simply involves looking up the appropriate normal probability density. As regards  $P(\boldsymbol{\theta} | \Pi, \mathbf{y})$ , since the prior is conjugate to the likelihood of the observations given the partition, we can sample the model parameters easily, first drawing a covariance matrix and then drawing a mean given the covariance matrix.

### Evaluating samples from the posterior density on partitions and category parameters

Each Gibbs sample contains an estimate of both the model parameters and of the partition. The question now arises of how to use these samples, and how to assess the overall quality of the posterior distribution. In terms of assessing model parameters, it is not really appropriate in principle to average across samples because our problem is PERMUTATION INVARIANT in the categories: if we took one set of model parameters  $\boldsymbol{\theta}$  fit and then permuted the category labels to get another set of parameters  $\boldsymbol{\theta}'$ , there would be no reason to prefer one or the other. For this reason, it is common simply to get a sense of the posterior distribution over model parameters by inspecting a single sample (or by inspecting a small number of samples individually). In Figure 9.6, we have taken the last of 1000 samples and plotted its characteristic ellipses alongside the “true” characteristic ellipses for the four categories’ Gaussian distributions (to be precise, the bias-corrected maximum-likelihood estimated Gaussians based on true category membership of our observations). Overall, the unsupervised model has done quite a good job of finding categories that turn out to match the underlying categories fairly well.

We now turn to evaluating the quality of the partitions themselves. This task is made a bit tricky by the fact that there are no deep principles available to map categories learned by our model to “true” categories given in the data. That is, suppose that we draw a sample partition from the model which puts observations  $y_1, y_2$ , and  $y_3$  in category 1, and observations  $y_4$  and  $y_5$  in category 2; but that underlyingly,  $y_1$  and  $y_2$  are instances of the vowel [i], and  $y_3, y_4$ , and  $y_5$  are instances of the vowel [ɪ]. How accurate is the partition? A common approach is to choose the MAPPING from model categories to true categories that maximizes the proportion of correctly assigned observations, and use this proportion as a measure of model performance; in our toy example, we would assign category 1 to [i] and category 2

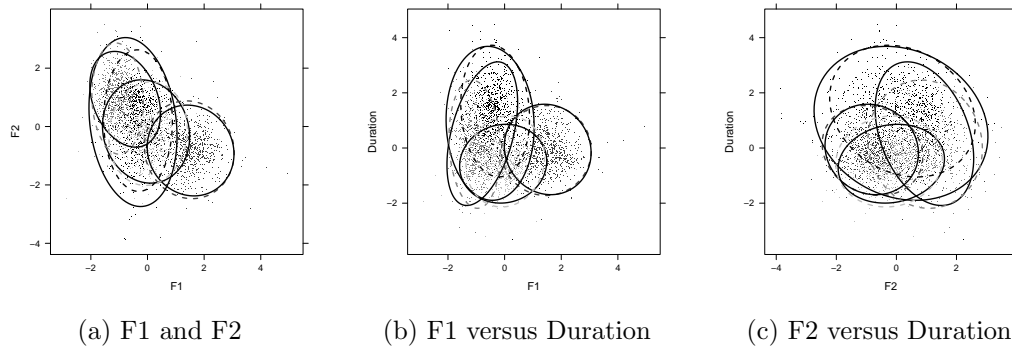


Figure 9.6: Categories inferred through unsupervised learning (solid lines) compared with the categories estimated from the same data with known categories via bias-corrected maximum likelihood (dotted lines)

to [i], and assess model performance at 80%. This approach has the disadvantage that the number of possible mappings increases exponentially with the number of true and learned categories, and greedy algorithms must be used to approximate the best mapping. Additionally, the mapping approach can become both conceptually and practically problematic when the number of latent categories is not equal to the number of true categories.<sup>2</sup>

Here we cover two other methods of assessing model performance in assignment of categories to partitions. The first method is purely visual: for a single sample partition, we construct a CONFUSION TABLE between true and guessed categories. Such a confusion table is in Figure 9.7a. In this table, each column is a model category and each row is a true vowel. The size of the square in each cell indicates the proportion of the guessed category that actually belongs to the true category. In this case, the model has some trouble distinguishing [i] from [e], but otherwise does a good job in distinguishing among the vowels. The corresponding confusion table for a supervised version of the model can be seen in Figure 9.7b; it is very clean.

The last method we cover here is assessing how well the model performs on average at the decision of whether two randomly drawn observations fall in the same category. This method is free of the matching-accuracy measure’s need for a greedy cluster-assignment algorithm, but has similar problems to matching accuracy when the number of categories in the model is not equal to the number of true categories. Like the matching-accuracy measure and unlike the confusion-matrix approach, the same-category measure can take advantage of averaging across multiple samples. In our case, drawing 1000 observation pairs without replacement from our dataset and averaging over every tenth sample from among the final 100 samples gives us a same-category measure of 0.7. The same measure can be applied to

<sup>2</sup>For example, one has to choose whether to impose a one-to-one constraint on the mapping between latent and true categories. When there are fewer true categories than there are latent categories, imposing a one-to-one constraint penalizes the model for making use of its full category inventory. On the other hand, if a one-to-one constraint is not imposed, then in the logical extreme case when the model has as many categories as it has observations, it is guaranteed to be assessed perfect accuracy!



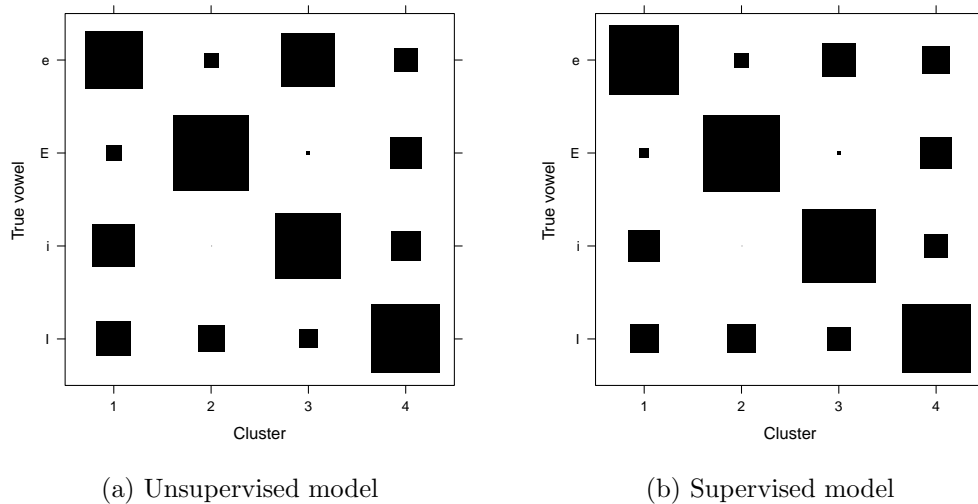


Figure 9.7: Confusion tables in the four-vowel unsupervised learning model.

the supervised model, with a performance here of 0.744. In our case, a baseline measure using random-category assignment would have performance of  $0.75^2 + 0.25^2 = 0.625$ : there is a 75% chance that the observations will be randomly drawn from different categories and the random category-assigner will assign different categories, and a 25% chance that the observations will be randomly drawn from the same category and the random assigner will assign the same category. Thus we can see that by this measure that the unsupervised learning model has learned enough to discriminate with accuracy far greater than chance whether two observations belong to the same category, but (unsurprisingly) has not reached the performance of the supervised learner.

### 9.1.2 Learning mixture probabilities

The simulation we set up was made artificially easy by the equal representation of all four vowels in our learning data, together with our hard prior constraint that category probabilities  $\phi_i$  be identical. In real learning scenarios, underlying categories occur with different frequencies—for example, [ɪ] occurs more frequently in spoken American English than the other vowels here [add ref]. In principle, one might suppose that these category frequencies might not be learned initially; once the categories themselves have been learned and become accurately identifiable in the linguistic input, their frequencies can be estimated relatively trivially. On the other hand, it is also possible that categories and their frequencies are learned simultaneously—indeed, this approach might facilitate learning of the categories themselves.

In our current framework, this approach would entail that inferences about the multinomial category probabilities  $\phi$  are drawn together with inferences about possible partitions  $\Pi$  and category structure parameters  $\theta$ . It is straightforward to accommodate this additional

aspect of inference within the Bayesian framework, putting a prior on  $\phi$ , as in Figure 9.3c. A natural prior to use is the Dirichlet distribution, as it is conjugate to the multinomial (Sections XXX and B.8):

$$\phi \sim \mathcal{D}(\Sigma_\phi)$$

This combines with the other probability distributions we have used earlier:

$$\begin{aligned} \Sigma_i &\sim \mathcal{IW}(\Sigma_0, \nu) \\ \mu_i | \Sigma &\sim \mathcal{N}(\mu_0, \Sigma_i/A) \\ i &\sim \text{Multinom}(\phi) \\ y &\sim \mathcal{N}(\mu_i, \Sigma_i) \end{aligned}$$

to form our complete model. As discussed [elsewhere], the Dirichlet parameters  $\Sigma_\phi$  are a vector of “pseudocounts”, with larger total summed pseudocounts indicating higher confidence in the underlying probability. For our four-category Dirichlet prior, we use  $\Sigma_\phi = \langle 4, 4, 4, 4 \rangle$ , giving us a prior on category probabilities biased toward but not enforcing a uniform distribution. As before, we can use Gibbs sampling to draw from the posterior distribution over category probabilities; this is easy because of the conjugacy of the Dirichlet distribution to the multinomial distribution, which gives us an analytic form for  $P(\phi | \Sigma_\phi, \Pi)$ .

We evaluate the performance of this model by changing the probability of each vowel based loosely the frequencies in American English, but exaggerating the skew across vowels to emphasize the consequences for learning. We assign [e] a probability of 0.04, [ɛ] a probability of 0.05, [i] a probability of a probability of 0.29, and [ɪ] a probability of 0.62 respectively. The inability to learn category frequencies becomes an especially prominent issue with smaller samples (see Exercise ?? for more on this), so we give only 30 per speaker, distributed according to the probabilities just given. Results from the final of 1000 Gibbs samples are plotted in Figure 9.8. The learning problem has become much more difficult, due both to the smaller amount of data provided and to the less-constrained model space. Nevertheless, there are clear signs that the model has learned the qualitative structure of the four vowel categories.

These results can be contrasted with those obtained when our previous model—in which  $\phi_i$  were constrained to be equiprobable—is applied to these same data. The results from the final of 1000 Gibbs samples are shown in Figure 9.9. This model has also made some progress in the learning of category structure; there is a crucial difference, however, in that there is much heavier overlap across categories in the F1 and F2 dimensions. The difference in the results of the two models becomes more apparent when confusion matrices are constructed, as in Figure 9.10. The category-frequency learner clearly has trouble distinguishing between [e] and [i], but it has succeeded in learning a three-way distinction between [e]/[i], [ɛ], and [ɪ], and it has learned that [ɪ] is the most frequent category. The fixed-category-frequency learner, on the other hand, has not as clearly learned the difference between [ɛ] and [ɪ], and

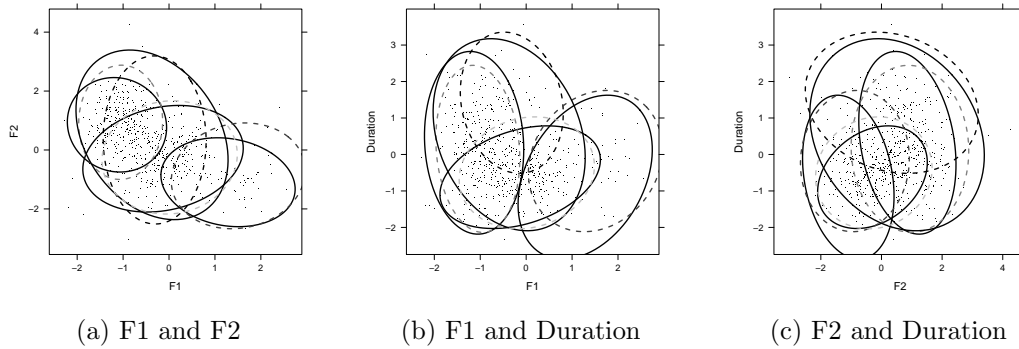


Figure 9.8: Learning category probabilities too.

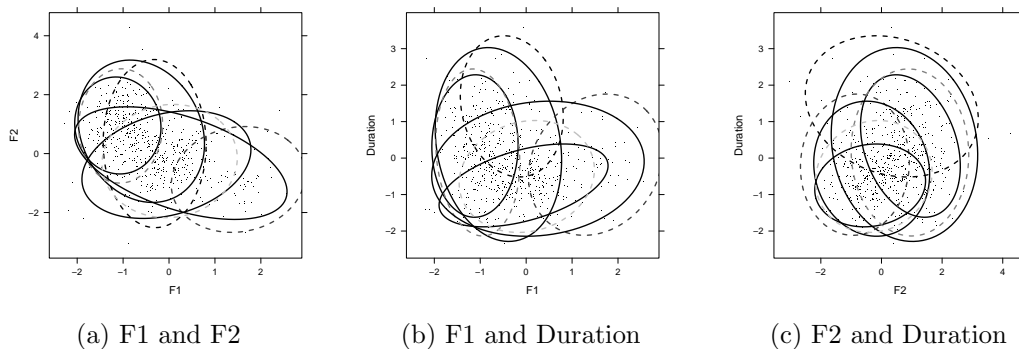
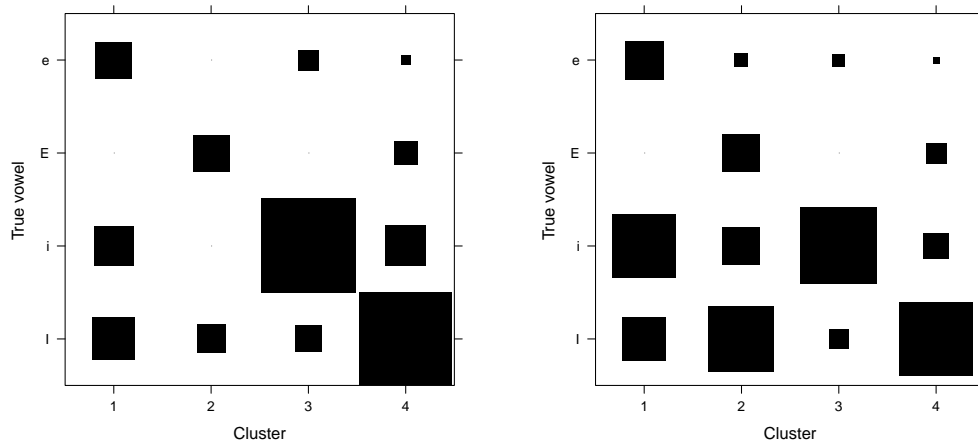


Figure 9.9: Inferences in the skewed category-frequency case without the ability to learn category probability.

underlying [i] observations are scattered widely across its four latent categories. \*\*\*Same-different performance: catfreqlearn: 0.627; fixed: 0.625

## 9.2 Latent Dirichlet Allocation and Topic Models

Another prominent dimension of language where the acquisition of latent structure is widely considered to be of great importance is *word meaning*. The sheer rate of word learning in native speakers is staggering: current estimates are that American native English speakers know an average of 40–100,000 words by age 17, and are thus learning new words at a rate of somewhere around ten a day (Nagy and Anderson, 1984; Landauer, 1986; Bloom, 2001). Researchers such as Landauer and Dumais (1997) have argued that most of the learning of word meanings must be implicit, since there is relatively little explicit instruction about the meanings of new words in classrooms, children’s and young adults’ reading materials, or in daily conversational interaction. The strong implication is that native speakers are able to construct sufficiently constraining representations of the possible meanings that an unknown word may have on the basis of the context (both linguistic and extra-linguistic) in which



(a) With learning of category frequencies (b) Without learning of category frequencies

Figure 9.10: Confusion matrices in the skewed category-frequency case with and without the ability to learn category frequencies

it appears to greatly facilitate learning of the unknown word’s meaning. Investigation into the nature of these constraining representations of context has been an area of considerable interest in linguistics, cognitive science, and natural language processing.

One of the most extensively developed lines of research in this area has treated the representation of linguistic context with the use of VECTOR SPACES. On the most straightforward interpretation of this approach, linguistic meaning can be thought of as a continuous multi-dimensional semantic space populated by words meanings, each of which can be represented by a real-valued vector. The implications for how word learning could be facilitated by such representations are clear: if the learner is ultimately exposed to  $N$  words but the underlying vector space is of dimension  $T \ll N$ , the majority of the learning task may simply be learning the underlying vector space, which could be accomplished by exposure to a much smaller number of words  $N'$  on the order of  $T$ ; after learning the vector space, fitting in the meanings of the remaining  $N - N'$  words would be a much easier task.

Although it seems fairly clear that vector spaces are far from adequate as a complete model of word meaning—they leave unaddressed important qualitative facets of lexical semantics such as argument structure, aspect, distinctions between different types of antonymy, scalarity of adjectives, and so forth—they have proven quite popular as a research tool, in no small part because formal methods for inferring vector-space meaning representations have been able to draw upon well-developed statistical techniques. Here, we cover two of the leading approaches to learning vector-space representations of word meaning: LATENT SEMANTIC ANALYSIS (LSA; Landauer and Dumais, 1997) and LATENT DIRICHLET ALLOCATION (LDA, also known as TOPICS MODELS; Blei et al., 2003; Griffiths and Steyvers, 2004; Griffiths et al., 2007). Viewed from the framework we have been pursuing in this textbook, LSA can be seen as a pure likelihood-based model based on assumptions of underlying nor-

mally distributed observations; LDA as most commonly practiced can be seen as a Bayesian model based on assumptions of multinomial-distributed observations. Both can be seen as instances of DIMENSIONALITY-REDUCTION techniques, where a lossless representation of the distribution of  $N$  word types across  $D$  documents is compressed down to a lossy representation in  $T \ll N$  dimensions of word meaning. We cover LSA in Section XXX; here we cover LDA.

### 9.2.1 Formal specification

Formally, an LDA model involves the following components:

- A fixed vocabulary  $V$  of words;
- A set of  $T$  topics;
- For the  $i$ -th topic  $z_i$ , a multinomial distribution with parameters  $\phi_i$  over all words in the vocabulary;
- For each document  $d$ , a multinomial distribution with parameters  $\theta_d$  over the topics—this distribution can be called the TOPIC MIXTURE for  $d$ ;
- A distribution over topic mixtures, generally taken to be a Dirichlet distribution with parameters  $\sigma_\theta$ ;
- Some distribution  $\zeta$  over the number of words contained in a document.

The generative process for each document  $d$  is as follows: first, the number of words in  $d$  is drawn from  $\zeta$ . Then, each word is generated by first sampling a topic  $z_j$  from  $\theta_d$ , and then sampling a word  $w_j$  from  $\phi_j$ . This process is used independently for each document. LDA as a graphical model is shown in Figure 9.11a. In the notation we've been using for specifying the individual component probability distributions in the graphical model, we have

$$\begin{aligned}\theta &\sim \mathcal{D}(\sigma_\theta) \\ z_i|\theta &\sim \text{Multinom}(\theta) \\ w_i|z_i, \phi_i &\sim \text{Multinom}(\phi_i)\end{aligned}$$

At this point, careful comparison with Figure ?? is worthwhile. LDA is a hierarchical model in which the individual observations (words) are grouped into clusters (documents), and each cluster has specific properties that are drawn from some overall distribution governing inter-cluster variability. Additionally, in LDA, as with hierarchical models in Figure ??, some parameters are shared across all clusters. The key difference in the structure of an LDA model as compared with the hierarchical models covered in Chapter 8 is that in LDA, the

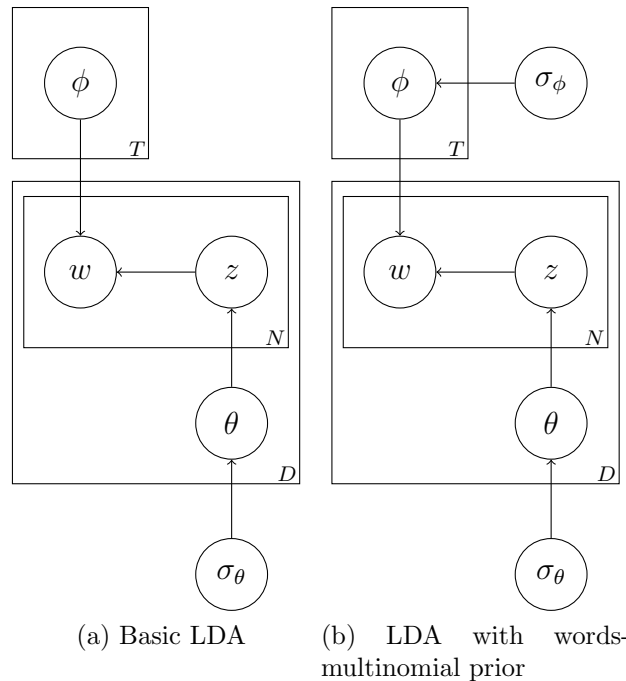


Figure 9.11: Latent Dirichlet Allocation as a generative model of words in documents

influence of the document-specific parameters  $\theta$  is mediated completely through the assignment of topic identity for each word. That is, document-specific properties  $\theta$  are conditionally independent of word identity given topic identity and the topic-specific word distributions  $\phi$ . In the hierarchical models of Chapter 8, in contrast, cluster-specific properties directly affected the probability distribution on the response.

3

- state distr's used for LDA
- comment on what's done and what else might be considered

### Collapsed Gibbs Sampling

The techniques available for inspecting the posterior distribution over partitions (topic assignments) ( $z$ ), document-specific topic mixtures ( $\theta$ ), and topic-specific word distributions ( $\phi$ ) in Latent Dirichlet Allocation provide a particularly clear example of how efficiencies in Bayesian inference for relatively complex graphical models can sometimes be found with appropriate choices of prior distribution. Consulting Figure 9.11b, we see that nodes in the graph which are either observed or set by researchers are  $\sigma_\phi$ ,  $\sigma_\theta$ , and  $w$ . In a typical Gibbs sampling approach, we would alternate between sampling values at the  $\phi$ ,  $\theta$ , and  $z$  nodes

<sup>3</sup>The tradition within the topic models literature is to use a slightly different nomenclature for cluster parameters and distribution on cluster identities, but we'll stick with the current nomenclature for consistency.

from the appropriate conditional probability distributions. Due to the use of Dirichlet distributions as priors and their conjugacy to the multinomial, all the distributions in question for the LDA Gibbs sampler turn out to be either multinomial or Dirichlet, this approach is perfectly feasible.

However, it turns out to be possible to do even better than this. Let us imagine that we know topics for all the words in our dataset except for the  $j$ th word of the  $i$ th document, which we can call  $w_{ij}$ . We can denote this set of topic assignments as  $\mathbf{z}-ij$ . The distribution on the topic for that word,  $z_{ij}$ , conditional on  $\mathbf{z}-ij$  can be written using Bayes' rule as follows:

$$P(z_{ij}|\mathbf{w}, \mathbf{z}-ij, \sigma_\theta, \sigma_\phi) = \frac{P(w_{ij}|\mathbf{z}\mathbf{w}_{-ij}, \sigma_\theta, \sigma_\phi)P(z_{ij}|\mathbf{w}_{-ij}, \mathbf{z}-ij, \sigma_\theta, \sigma_\phi)}{P(w_{ij}|\mathbf{z}-ij, \mathbf{w}_{-ij}, \sigma_\theta, \sigma_\phi)} \quad (9.7)$$

$$(9.8)$$

Let us ignore the denominator (which is just a normalizing constant) and focus on the numerator. Inspecting Figure 9.11b, we can see that the first term (the likelihood of  $w_{ij}$  given its topic assignment  $z_{ij}$ ) is conditionally independent of  $\theta$  and  $\sigma_\theta$ , since the only connection between these nodes and  $w_{ij}$  is strictly downstream in the graph and goes through  $z_{ij}$ , which is being conditioned on (see Chapter C). However,  $w_{ij}$  is *not* independent of  $\phi$ , which we do not know. In this case, we need to marginalize:

$$P(w_{ij}|\mathbf{z}\mathbf{w}_{-ij}, \sigma_\phi) = \int_{\phi} P(w_{ij}|\mathbf{z}\mathbf{w}_{-ij}, \phi)P(\phi|\mathbf{z}\sigma_\phi, \mathbf{w}_{-ij}) d\phi \quad (9.9)$$

Recall that  $\phi$  is a set of multinomial parameters—one multinomial for each topic. Since we are concerned only with the probability of a word generated from topic  $z_{ij}$ , we can ignore all the multinomial parameters except for those associated with this topic. Now, there are only two sources of information regarding these multinomial parameters: (i) the prior, and (ii) the words that have been observed to be emitted from topic  $z_{ij}$ . This prior is Dirichlet and the likelihood is multinomial, so that the predictive distribution over  $w_{ij}$  is a DIRICHLET-MULTINOMIAL model (Section B.8; compare with the beta-binomial model introduced in Section 4.4.2). For each word in the lexicon  $w$ , let the prior parameter for this word be denoted as  $\sigma_\phi(w)$  and the number of observations of  $w$  to which topic  $z_{ij}$  has been assigned in the current sample as  $c(w)$ . The Dirichlet-multinomial model gives a simple analytic form for the marginal probability of  $w_{ij}$ :

$$P(w_{ij}|\mathbf{z}\mathbf{w}_{-ij}, \sigma_\phi) = \frac{\sigma_\phi(w) + c(w)}{\sum_{w'} \sigma_\phi(w') + c(w')}$$

That is, we can *marginalize* over the parameters  $\phi$  instead of sampling from them.

If we look at the second term of the numerator in Equation (9.7), we can see that it can also be expressed as a marginalization over document-specific multinomial parameters  $\theta$ , with several conditioning variables being irrelevant:

$$P(z_{ij}|\mathbf{w}_{-ij}, \mathbf{z}_{-ij}, \sigma_\theta, \sigma_\phi) = \int_{\theta} P(z_{ij}|\theta)P(\theta|\mathbf{z}_{-ij}, \sigma_\theta) d\theta$$

In the same way as before, the first term in this integral—the likelihood—is multinomial, and the second term—the prior—is Dirichlet, giving us a Dirichlet-multinomial predictive distribution. If for each topic  $z$  we denote its prior parameter as  $\sigma_\theta(z)$  and the number of words within document  $i$  that in the current sample have been assigned to  $z$  as  $c_i(z)$ , this predictive distribution has a simple form:

$$P(z_{ij}|\mathbf{w}_{-ij}, \mathbf{z}_{-ij}, \sigma_\theta, \sigma_\phi) = \frac{\sigma_\theta(w) + c_i(w)}{\sum_{z'} \sigma_\theta(z') + c_i(z')}$$

Combining the two expressions we get

$$P(w_{ij}|\mathbf{z}\mathbf{w}_{-ij}, \sigma_\phi) \propto \frac{(\sigma_\phi(w) + c(w)) (\sigma_\theta(w) + c_i(w))}{(\sum_{w'} \sigma_\phi(w') + c(w')) (\sum_{z'} \sigma_\theta(z') + c_i(z'))}$$

which eliminates the need to sample any multinomial parameters whatsoever. This technique of marginalizing over some latent variables in a Gibbs sampler is known as the COLLAPSED GIBBS SAMPLER, because the samples collapse (marginalize) over some of the nodes in the graphical model. The collapsed Gibbs sampler is generally more efficient than the corresponding uncollapsed Gibbs sampler (Liu, 1994)



# Chapter 10

## Probabilistic Grammars

### 10.1 Outline

- HMMs
- PCFGs
- pTSGs and pTAGS
- Highlight: Zuidema et al., 2008, CogSci; Cohn & Goldwater & co., NAACL 2009

### 10.2 Weighted Finite-State Transducers

A WEIGHTED FINITE-STATE TRANSDUCER (wFSA)  $T$  has the following components:

- A finite INPUT ALPHABET  $\Sigma$ , which includes the empty string  $\epsilon$ ;
- A finite OUTPUT ALPHABET  $\Delta$ , which includes the empty string  $\epsilon$ ;
- A finite set of STATES  $q$ , with distinguished members as follows:
  - One element of  $q$  is the INITIAL STATE  $q_0$ ;
  - Some subset of  $q$  are FINAL STATES  $F$ , and each final state  $q_i$  has an associated weight  $w_i$ ;
- A set of TRANSITIONS  $E$ , with each transition consisting of:
  - A source state  $q$  and a destination state  $q'$ ;
  - A mapping of an input symbol  $\sigma \in \Sigma$  to an output symbol  $\delta \in \Delta$ ;
  - An WEIGHT  $w$  associated with the transition.

|     |          |         |     |     |          |   |
|-----|----------|---------|-----|-----|----------|---|
| 1   | <b>S</b> | →NP VP  | 1   | Det | →the     | $N = \{S, NP, Det, N', N, VP, V, PP, P\}$ |
| 1   | NP       | →Det N' | 0.5 | N   | →dog     | $V = \{the, dog, cat, near, growled\}$    |
| 0.8 | N'       | →N      | 0.5 | N   | →cat     | $S = S$                                   |
| 0.2 | N'       | →N' PP  | 1   | P   | →near    |   |
| 1   | PP       | →P NP   | 1   | V   | →growled |   |
| 1   | VP       | →V      |     |     |          |   |

Figure 10.1: A simple probabilistic context-free grammar, with the start symbol in bold.

### 10.3 Probabilistic context-free grammars

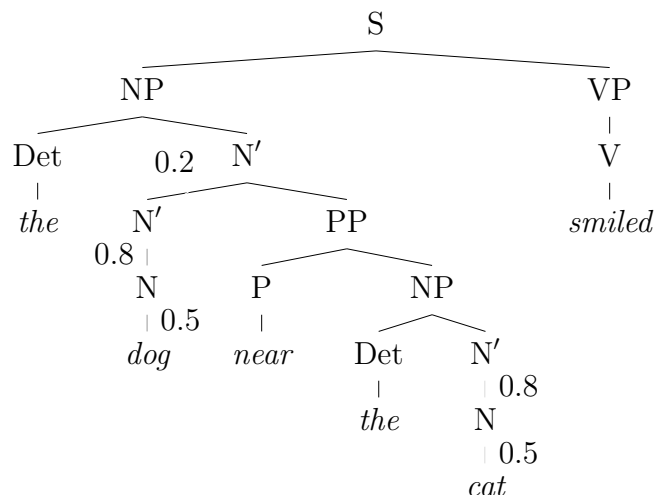
A probabilistic context-free grammar consists of a tuple  $(N, V, S, R, P)$  such that:

- $N$  is a finite set of non-terminal symbols;
- $V$  is a finite set of terminal symbols;
- $S$  is the start symbol;
- $R$  is a finite set of rules of the form  $X \rightarrow \alpha$  where  $X \in N$  and  $\alpha$  is a sequence of symbols drawn from  $V \cup N$ ;
- $P$  is a mapping from  $R$  into probabilities, such that for each  $X \in N$ ,

$$\sum_{[X \rightarrow \alpha] \in R} P(X \rightarrow \alpha) = 1$$

Figure 10.1 shows a simple PCFG. Figure 10.2 shows one of the trees that this PCFG generates. Note, however, that there are infinitely many trees that can be generated by even this simple grammar. Because the category PP can contain the category N', the rule  $N' \rightarrow N' PP$  effectively turns one N' into two. This rule can be applied arbitrarily many times. [consider a different example that involves center-embedding?]

- Use of a given PCFG
  - A PCFG gives probabilities of *trees*
  - The *yield* of a tree is (deterministically) a string
  - Thus we can use a given PCFG for all sorts of Bayesian inferences:
    - \*  $P(w)$
    - \*  $P(T|w)$
    - \*  $P(T|I)$  for some input  $I$
    - \*  $P(w|I)$  for some input  $I$



$$\begin{aligned}
 P(T) &= 1 \times 1 \times 1 \times 0.2 \times 0.8 \times 0.5 \times 1 \times 1 \times 1 \times 1 \times 0.8 \times 0.5 \times 1 \times 1 \\
 &= 0.032
 \end{aligned}$$

Figure 10.2: A simple tree

\*  $P(w_i|w_{1..i-1})$  – Prediction (conditional word probabilities)

- Unifying these notions as probabilistic intersection
- This can be done efficiently when the evidence is formalizable as a weighted finite-state automaton/transducer

- Estimation of PCFGs

- Supervised
- Unsupervised: EM and Bayesian

### 10.3.1 Probabilities from wCFGs

[discuss wCFGs]

The complete structure generated by a CFG or CFG derivation is a *tree*. We have already discussed the case of PCFGs; now let us turn to the probabilistic interpretation of weighted CFGs (wCFGs). Define a tree  $T$ 's WEIGHT  $w(T)$  assigned by the wCFG as the product of the weights of each of the rules that make up the tree (with a rule having its weight counted as many times as it is used). Now consider the infinite set  $\mathcal{T}$  of all trees that can be generated by the PCFG. The normalizing constant  $Z$  (see Section 2.8) is just the sum of the weights of all these trees:

$$Z \equiv \sum_{T \in \mathcal{T}} w(T)$$

The probability assigned to tree  $T$  is then just its weight divided by this normalizing constant:<sup>1</sup>

$$P(T) = \frac{w(T)}{Z}$$

Following the definitions given back in Chapter ??, if the normalizing constant  $Z$  is equal to 1, then the rule weights alone give us a PROPER probability distribution over the trees, with a tree’s probability equal to the product of its rule weights.

Notice, though, that nothing we have said thus far entails anything about bounds on the weights of individual rules in wCFGs. For example, the following wCFG (with start symbol NP) is proper by the definitions given above, even though two of the rule weights exceed 1:

$$\begin{array}{ll} 1/4 & \text{NP} \rightarrow \text{N} & 2 & \text{N} \rightarrow \text{cats} \\ & & 2 & \text{N} \rightarrow \text{dogs} \end{array}$$

One might then wonder whether wCFGs and PCFGs are equivalent formalisms for putting probability distributions over tree structures. Fortunately, the answer is yes: for every wCFG with finite normalization constant there is a PCFG that puts the same probability distribution on possible trees (Abney et al., 1999; Chi, 1999).<sup>2</sup> Furthermore, there are constructive methods for finding this PCFG (Chi, 1999). This close equivalence between wCFGs and PCFGs will come in very convenient in the later sections of this chapter.

## 10.4 Exercises

### Exercise 10.1

Estimating a small PCFG

Consider the so-called “NP/S” garden path in online language comprehension:

- (1)    a.    Sammy heard the noise bothered Pat.  
       b.    Sammy heard that the noise bothered Pat.

Write a small context-free grammar with enough rules to cover this data, including the local ambiguity in which *the noise* is misparsed as the direct object of *heard*. Estimate rule

<sup>1</sup>Provided the normalizing constant  $Z$  is finite. If  $Z$  is not finite, then the probability of a tree is undefined.

<sup>2</sup>It has additionally been shown that the conditional distribution on trees given strings— $P(T|w)$ —determined by a given wCFG can also be expressed by some PCFG, even when the overall partition function of the wCFG is non-finite (Smith and Johnson, 2007).

probabilities as follows: for each rule, write a tree-search expression corresponding using a `tgrep2`, `Tregex`, or `TIGERSearch` that obtains a frequency count from a parsed corpus such as the Penn Treebank. Use relative frequency estimation or some other estimation technique to obtain rule probabilities from these counts. Then, based on your resulting PCFG, compute the

- posterior probabilities of the direct-object and sentential-complement analyses of the sentences for each sentence prefix that contains at least the input tokens through *the*.
- Conditional word negative log-probabilities (i.e. surprisal values) of each word in each sentence beginning with *the*.

Finally, change something in your grammar (up to you). How does the change you make affect the above probabilities? Discuss.

# Appendix A

## Mathematics notation and review

This appendix gives brief coverage of the mathematical notation and concepts that you'll encounter in this book. In the space of a few pages it is of course impossible to do justice to topics such as integration and matrix algebra. Readers interested in strengthening their fundamentals in these areas are encouraged to consult XXX [calculus] and Healy (2000).

### A.1 Sets ( $\{\}$ , $\cup$ , $\cap$ , $\emptyset$ )

The notation  $\{a, b, c\}$  should be read as “the set containing the elements  $a$ ,  $b$ , and  $c$ ”. With sets, it's sometimes a convention that lower-case letters are used as names for elements, and upper-case letters as names for sets, though this is a weak convention (after all, sets can contain anything—even other sets!).

$A \cup B$  is read as “the union of  $A$  and  $B$ ”, and its value is the set containing exactly those elements that are present in  $A$ , in  $B$ , or in both.

$A \cap B$  is read as “the intersection of  $A$  and  $B$ ”, and its value is the set containing only those elements present in both  $A$  and  $B$ .

$\emptyset$ , or equivalently  $\{\}$ , denotes the empty set—the set containing nothing. Note that  $\{\emptyset\}$  isn't the empty set—it's the set containing only the empty set, and since it contains something, it isn't empty!

[introduce set complementation if necessary]

#### A.1.1 Countability of sets

[briefly describe]

### A.2 Summation ( $\sum$ )

Many times we'll want to express a complex sum of systematically related parts, such as  $1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5}$  or  $x_1 + x_2 + x_3 + x_4 + x_5$ , more compactly. We use SUMMATION notation for this:

$$\sum_{i=1}^5 \frac{1}{i} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} \qquad \sum_{i=1}^5 x_i = x_1 + x_2 + x_3 + x_4 + x_5$$

In these cases,  $i$  is sometimes called an INDEX VARIABLE, linking the RANGE of the sum (1 to 5 in both of these cases) to its contents. Sums can be nested:

$$\sum_{i=1}^2 \sum_{j=1}^2 x_{ij} = x_{11} + x_{12} + x_{21} + x_{22} \qquad \sum_{i=1}^3 \sum_{j=1}^i x_{ij} = x_{11} + x_{21} + x_{22} + x_{31} + x_{32} + x_{33}$$

Sums can also be infinite:

$$\sum_{i=1}^{\infty} \frac{1}{i^2} = \frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \dots$$

Frequently, the range of the sum can be understood from context, and will be left out; or we want to be vague about the precise range of the sum. For example, suppose that there are  $n$  variables,  $x_1$  through  $x_n$ . In order to say that the sum of all  $n$  variables is equal to 1, we might simply write

$$\sum_i x_i = 1$$

### A.3 Product of a sequence ( $\prod$ )

Just as we often want to express a complex sum of systematically related parts, we often want to express a product of systematically related parts as well. We use PRODUCT notation to do this:

$$\prod_{i=1}^5 \frac{1}{i} = 1 \times \frac{1}{2} \times \frac{1}{3} \times \frac{1}{4} \times \frac{1}{5} \qquad \prod_{i=1}^5 x_i = x_1 x_2 x_3 x_4 x_5$$

Usage of product notation is completely analogous to summation notation as described in Section A.2.

### A.4 "Cases" notation ( $\{ \}$ )

Some types of equations, especially those describing probability functions, are often best expressed in the form of one or more conditional statements. As an example, consider a six-sided die that is weighted such that when it is rolled, 50% of the time the outcome is

a six, with the other five outcomes all being equally likely (i.e. 10% each). If we define a discrete random variable  $X$  representing the outcome of a roll of this die, then the clearest way of specifying the probability mass function for  $X$  is by splitting up the real numbers into three groups, such that all numbers in a given group are equally probable: (a) 6 has probability 0.5; (b) 1, 2, 3, 4, and 5 each have probability 0.1; (c) all other numbers have probability zero. Groupings of this type are often expressed using “cases” notation in an equation, with each of the cases expressed on a different row:

$$P(X = x) = \begin{cases} 0.5 & x = 6 \\ 0.1 & x \in \{1, 2, 3, 4, 5\} \\ 0 & \text{otherwise} \end{cases}$$

## A.5 Integration ( $\int$ )

Sums are always over countable (finite or countably infinite) sets. The analogue over a continuum is INTEGRATION. Correspondingly, you need to know a bit about integration in order to understand continuous random variables. In particular, a basic grasp of integration is essential to understanding how Bayesian statistical inference works.

One simple view of integration is as computing “area under the curve”. In the case of integrating a function  $f$  over some range  $[a, b]$  of a one-dimensional variable  $x$  in which  $f(x) > 0$ , this view is literally correct. Imagine plotting the curve  $f(x)$  against  $x$ , extending straight lines from points  $a$  and  $b$  on the  $x$ -axis up to the curve, and then laying the plot down on a table. The area on the table enclosed on four sides by the curve, the  $x$ -axis, and the two additional straight lines is the integral

$$\int_a^b f(x) dx$$

This is depicted graphically in Figure A.1a.

The situation is perhaps slightly less intuitive, but really no more complicated, when  $f(x)$  crosses the  $x$ -axis. In this case, area under the  $x$ -axis counts as “negative” area. An example is given in Figure A.1b; the function here is  $f(x) = \frac{1}{2}(2.5 - x)$ . Since the area of a triangle with height  $h$  and length  $l$  is  $\frac{lh}{2}$ , we can compute the integral in this case by subtracting the area of the smaller triangle from the larger triangle:

$$\int_1^3 f(x) dx = \frac{1.5 \times 0.75}{2} - \frac{0.5 \times 0.25}{2} = 0.5$$

Integration also generalizes to multiple dimensions. For instance, the integral of a function  $f$  over an area in two dimensions  $x$  and  $y$ , where  $f(x, y) > 0$ , can be thought of as the volume



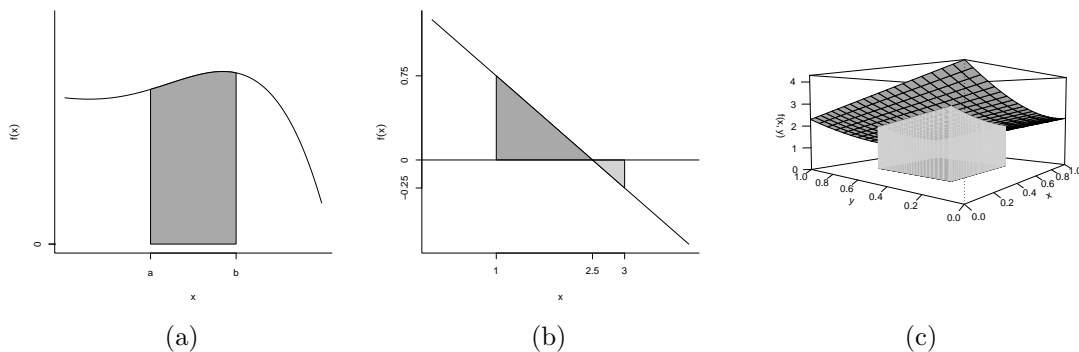


Figure A.1: Integration

enclosed by projecting the area's boundary from the  $x, y$  plane up to the  $f(x, y)$  surface. A specific example is depicted in Figure A.1c, where the area in this case is the square bounded by  $1/4$  and  $3/4$  in both the  $x$  and  $y$  directions.

$$\int_{\frac{1}{4}}^{\frac{3}{4}} \int_{\frac{1}{4}}^{\frac{3}{4}} f(x, y) dx dy$$

An integral can also be over the *entire* range of a variable or set of variables. For instance, one would write an integral over the entire range of  $x$  as  $\int_{-\infty}^{\infty} f(x) dx$ . Finally, in this book and in the literature on probabilistic inference you will see the abbreviated notation  $\int_{\theta} f(\theta) d\theta$ , where  $\theta$  is typically an ensemble (collection) of variables. In this book, the proper interpretation of this notation is as the integral over the entire range of all variables in the ensemble  $\theta$ .

### A.5.1 Analytic integration tricks

Computing an integral ANALYTICALLY means finding an exact form for the value of the integral. There are entire books devoted to analytic integration, but for the contents of this book you'll get pretty far with just a few tricks.

1. **Multiplication by constants.** The integral of a function times a constant  $C$  is the product of the constant and the integral of the function:

$$\int_a^b C f(x) dx = C \int_a^b f(x) dx$$

2. **Sum rule.** The integral of a sum is the sum of the integrals of the parts:

$$\int_a^b [f(x) + g(x)] dx = \int_a^b f(x) dx + \int_a^b g(x) dx$$

3. **Expressing one integral as the difference between two other integrals:** For  $c < a, b$ ,

$$\int_a^b f(x) dx = \int_c^b f(x) dx - \int_c^a f(x) dx$$

This is an extremely important technique when asking whether the outcome of a continuous random variable falls within a range  $[a, b]$ , because it allows you to answer this question in terms of cumulative distribution functions (Section 2.6); in these cases you'll choose  $c = -\infty$ .

4. **Polynomials.** For any  $n \neq -1$ :

$$\int_a^b x^n dx = \frac{1}{n+1}(b^{n+1} - a^{n+1})$$

And the special case for  $n = -1$  is:

$$\int_a^b x^{-1} dx = \log b - \log a$$

Note that this generalization holds for  $n = 0$ , so that integration of a constant is easy:

$$\int_a^b C dx = C(b - a)$$

5. **Normalizing constants.** If the function inside an integral looks the same as the probability density function for a known probability distribution, then its value is related to normalizing constant of the probability distribution. [Examples: normal distribution; beta distribution; others?] For example, consider the integral

$$\int_{-\infty}^{\infty} \exp\left[-\frac{x^2}{18}\right] dx$$

This may look hopelessly complicated, but by comparison with Equation 2.20 in Section 2.10 you will see that it looks just like the probability density function of a normally distributed random variable with mean  $\mu = 0$  and variance  $\sigma^2 = 9$ , except that

it doesn't have the normalizing constant  $\frac{1}{\sqrt{2\pi\sigma^2}}$ . In order to determine the value of this integral, we can start by noting that any probability density function integrates to 1:

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx = 1$$

Substituting in  $\mu = 0, \sigma^2 = 9$  we get

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{18\pi}} \exp\left[-\frac{x^2}{18}\right] dx = 1$$

By the rule of multiplication by constants we get

$$\frac{1}{\sqrt{18\pi}} \int_{-\infty}^{\infty} \exp\left[-\frac{x^2}{18}\right] dx = 1$$

or equivalently

$$\int_{-\infty}^{\infty} \exp\left[-\frac{x^2}{18}\right] dx = \sqrt{18\pi}$$

giving us the solution to the original problem.

## A.5.2 Numeric integration

The alternative to analytic integration is NUMERIC integration, which means approximating the value of an integral by explicit numeric computation. There are many ways to do this—one common way is by breaking up the range of integration into many small pieces, approximating the size of each piece, and summing the approximate sizes. A graphical example of how this might be done is shown in Figure A.2, where each piece of the area under the curve is approximated as a rectangle whose height is the average of the distances from the  $x$ -axis to the curve at the left and right edges of the rectangle. There are many techniques for numeric integration, and we shall have occasional use for some of them in this book.

## A.6 Precedence ( $\prec$ )

The  $\prec$  operator is used occasionally in this book to denote LINEAR PRECEDENCE. In the syntax of English, for example, the information that “a verb phrase (VP) can consist of a verb (V) followed by a noun phrase (NP) object” is most often written as:

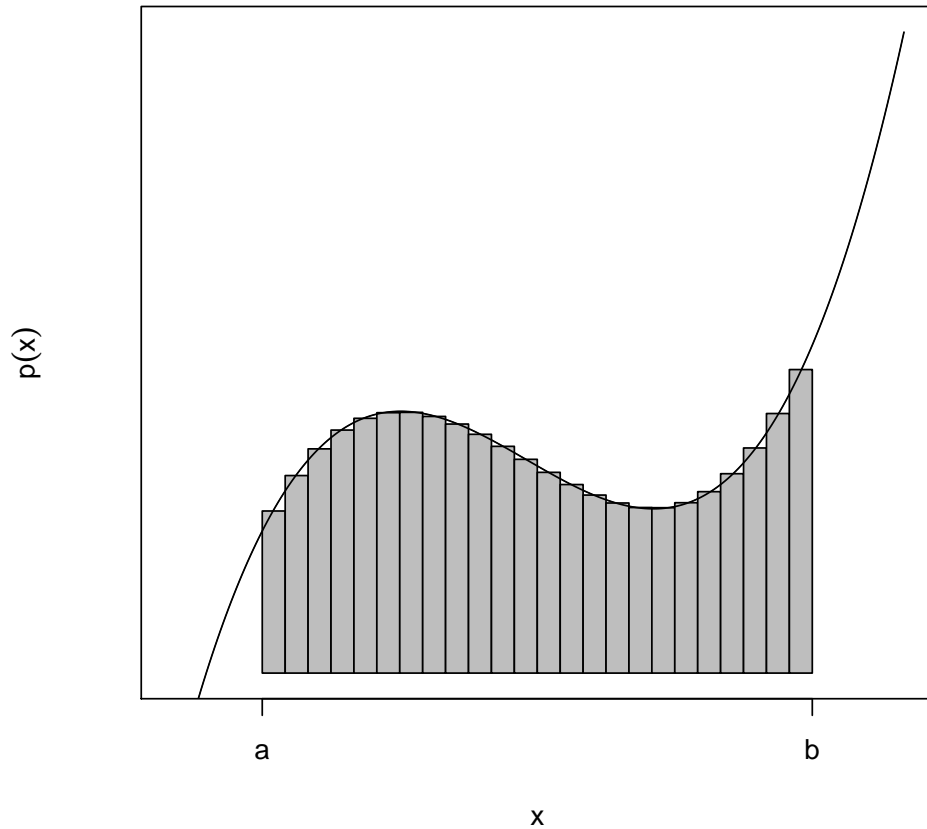


Figure A.2: Numeric integration

$$V \rightarrow V \text{ NP}$$

This statement combines two pieces of information: (1) a VP can be comprised of a V and an NP; and (2) in the VP, the V should precede the NP. In a syntactic tradition stemming from Generalized Phrase Structure Grammar (Gazdar et al., 1985), these pieces of information can be separated:

$$(1) V \rightarrow V, \text{ NP}$$

$$(2) V \prec \text{NP}$$

where  $V, \text{NP}$  means “the unordered set of categories V and NP”, and  $V \prec \text{NP}$  reads as “V precedes NP”.

## A.7 Combinatorics $\binom{n}{r}$

The notation  $\binom{n}{r}$  is read as “ $n$  choose  $r$ ” and is defined as the number of possible ways of selecting  $r$  elements from a larger collection of  $n$  elements, allowing each element to be chosen a maximum of once and ignoring order of selection. The following equality holds generally:

$$\binom{n}{r} = \frac{n!}{r!(n-r)!} \quad (\text{A.1})$$

The solution to the closely related problem of creating  $m$  classes from  $n$  elements by selecting  $r_i$  for the  $i$ -th class and discarding the leftover elements is written as  $\binom{n}{r_1 \dots r_m}$  and its value is

$$\binom{n}{r_1 \dots r_m} = \frac{n!}{r_1! \dots r_m!} \quad (\text{A.2})$$

Terms of this form appear in this book in the binomial and multinomial probability mass functions, and as normalizing constant for the beta and Dirichlet distributions.

## A.8 Basic matrix algebra

There are a number of situations in probabilistic modeling—many of which are covered in this book—where the computations needing to be performed can be simplified, both conceptually and notationally, by casting them in terms of MATRIX operations. A matrix  $\mathbf{X}$  of dimensions  $m \times n$  is a set of  $mn$  entries arranged rectangularly into  $m$  rows and  $n$  columns, with its entries indexed as  $x_{ij}$ :

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix}$$

For example, the matrix  $\mathbf{A} = \begin{bmatrix} 3 & 4 & -1 \\ 0 & -2 & 2 \end{bmatrix}$  has values  $a_{11} = 3$ ,  $a_{12} = 4$ ,  $a_{13} = -1$ ,  $a_{21} = 0$ ,  $a_{22} = -2$ , and  $a_{23} = 2$ . For a matrix  $\mathbf{X}$ , the entry  $x_{ij}$  is often called the  $i, j$ -th entry of  $\mathbf{X}$ .

If a matrix has the same number of rows and columns, it is often called a SQUARE matrix. A matrix of dimension  $m \times 1$ —that is, a single-column matrix—is often called a VECTOR.

**Diagonal and Identity matrices:** For a square matrix  $\mathbf{X}$ , the entries  $x_{ii}$ —that is, when the column and row numbers are the same—are called the DIAGONAL entries. A square matrix whose non-diagonal entries are all zero is called a DIAGONAL MATRIX. A diagonal matrix of size  $n \times n$  whose diagonal entries are all 1 is called the size- $n$  IDENTITY matrix. Hence  $\mathbf{A}$  below is a diagonal matrix, and  $\mathbf{B}$  below is the size-3 identity matrix.

$$\mathbf{A} = \begin{bmatrix} 3 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The  $n \times n$  identity matrix is sometimes notated as  $\mathbf{I}_n$ ; when the dimension is clear from context, sometimes the simpler notation  $\mathbf{I}$  is used.

**Transposition:** For any matrix  $\mathbf{X}$  of dimension  $m \times n$ , the TRANSPOSE of  $\mathbf{X}$ , or  $\mathbf{X}^T$ , is an  $n \times m$ -dimensional matrix such that the  $i, j$ -th entry of  $\mathbf{X}^T$  is the  $j, i$ -th entry of  $\mathbf{X}$ . For the matrix  $\mathbf{A}$  above, for example, we have

$$\mathbf{A}^T = \begin{bmatrix} 3 & 0 \\ 4 & -2 \\ -1 & 2 \end{bmatrix} \tag{A.3}$$

**Addition:** Matrices of like dimension can be added. If  $\mathbf{X}$  and  $\mathbf{Y}$  are both  $m \times n$  matrices, then  $\mathbf{X} + \mathbf{Y}$  is the  $m \times n$  matrix whose  $i, j$ -th entry is  $x_{ij} + y_{ij}$ . For example,

$$\begin{bmatrix} 3 & 0 \\ 4 & -2 \\ -1 & 2 \end{bmatrix} + \begin{bmatrix} -1 & -1 \\ 0 & 2 \\ 5 & 5 \end{bmatrix} = \begin{bmatrix} 2 & -1 \\ 4 & 0 \\ 4 & 7 \end{bmatrix} \tag{A.4}$$

**Multiplication:** If  $\mathbf{X}$  is an  $l \times m$  matrix and  $\mathbf{Y}$  is an  $m \times n$  matrix, then  $\mathbf{X}$  and  $\mathbf{Y}$  can be multiplied together; the resulting matrix  $\mathbf{XY}$  is an  $l \times n$  matrix. If  $\mathbf{Z} = \mathbf{XY}$ , the  $i, j$ -th entry of  $\mathbf{Z}$  is:

$$z_{ij} = \sum_{k=1}^m x_{ik}y_{kj}$$

For example, if  $A = \begin{bmatrix} 1 & 2 \\ -1 & 0 \\ 3 & 1 \end{bmatrix}$  and  $B = \begin{bmatrix} 3 & 4 & -1 & 6 \\ 0 & -5 & 2 & -2 \end{bmatrix}$ , we have

$$\begin{aligned} AB &= \begin{bmatrix} 1 \times 3 + 2 \times 0 & 1 \times 4 + 2 \times (-5) & 1 \times (-1) + 2 \times 2 & 1 \times 6 + 2 \times (-2) \\ (-1) \times 3 + 0 \times 0 & (-1) \times 4 + 0 \times (-5) & (-1) \times (-1) + 0 \times 2 & (-1) \times 6 + 0 \times (-2) \\ 3 \times 3 + 1 \times 0 & 3 \times 4 + 1 \times (-5) & 3 \times (-1) + 1 \times 2 & 3 \times 6 + 1 \times (-2) \end{bmatrix} \\ &= \begin{bmatrix} 3 & -6 & 3 & 2 \\ -3 & -4 & 1 & -6 \\ 9 & 7 & -1 & 16 \end{bmatrix} \end{aligned}$$

Unlike multiplication of scalars, matrix multiplication is **not** commutative—that is, it is not generally the case that  $\mathbf{XY} = \mathbf{YX}$ . In fact, being able to form the matrix product  $\mathbf{XY}$

does not even guarantee that we can do the multiplication in the opposite order and form the matrix product  $\mathbf{Y}\mathbf{X}$ ; the dimensions may not be right. (Such is the case for matrices  $A$  and  $B$  in our example.)

**Determinants.** For a square matrix  $\mathbf{X}$ , the DETERMINANT  $|\mathbf{X}|$  is a measure of the matrix’s “size”. In this book, determinants appear in coverage of the multivariate normal distribution (Section 3.5); the normalizing constant of the multivariate normal density includes the determinant of the covariance matrix. (The univariate normal density, introduced in Section 2.10, is a special case; there, it is simply the variance of the distribution that appears in the normalizing constant.) For small matrices, there are simple techniques for calculating determinants: as an example, the determinant of a  $2 \times 2$  matrix  $\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$  is  $|\mathbf{A}| = ad - bc$ . For larger matrices, computing determinants requires more general and complex techniques, which can be found in books on linear algebra such as Healy (2000).

**Matrix Inversion.** The INVERSE or RECIPROCAL of an  $n \times n$  square matrix  $\mathbf{X}$ , denoted  $\mathbf{X}^{-1}$ , is the  $n \times n$  matrix such that  $\mathbf{X}\mathbf{X}^{-1} = \mathbf{I}_n$ . As with scalars, the inverse of the inverse of a matrix  $\mathbf{X}$  is simply  $\mathbf{X}$ .

For example, the following pair of matrices are inverses of each other:

$$\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \qquad \mathbf{A}^{-1} = \begin{bmatrix} \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} \end{bmatrix}$$

Some useful properties of determinants and matrix inversion:

- The determinant of a matrix product is the product of the determinants:

$$|\mathbf{AB}| = |\mathbf{A}||\mathbf{B}|$$

- Because of this, the determinant of the inverse of a matrix is the reciprocal of the matrix’s determinant:

$$|\mathbf{A}^{-1}| = \frac{1}{|\mathbf{A}|}$$

## A.9 Miscellaneous notation

- $\propto$ : You’ll often see  $f(x) \propto g(x)$  for some functions  $f$  and  $g$  of  $x$ . This is to be read as “ $f(x)$  is proportional to  $g(x)$ ”, or “ $f(x)$  is equal to  $g(x)$  to within some constant”. Typically it’s used when  $f(x)$  is intended to be a probability, and  $g(x)$  is a function that obeys the first two axioms of probability theory, but is improper. This situation obtains quite often when, for example, conducting Bayesian inference.

# Appendix B

## More probability distributions and related mathematical constructs

This chapter covers probability distributions and related mathematical constructs that are referenced elsewhere in the book but aren't covered in detail.

### B.1 The gamma and beta functions

The GAMMA FUNCTION  $\Gamma(x)$ , defined for  $x > 0$ , can be thought of as a generalization of the factorial  $x!$ . It is defined as

$$\Gamma(x) = \int_0^{\infty} u^{x-1} e^{-u} du$$

and is available as a function in most statistical software packages such as R. The behavior of the gamma function is simpler than its form may suggest:  $\Gamma(1) = 1$ , and if  $x > 1$ ,  $\Gamma(x) = (x - 1)\Gamma(x - 1)$ . This means that if  $x$  is a positive integer, then  $\Gamma(x) = (x - 1)!$ .

The BETA FUNCTION  $B(\alpha_1, \alpha_2)$  is defined as a combination of gamma functions:

$$B(\alpha_1, \alpha_2) \stackrel{\text{def}}{=} \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1 + \alpha_2)}$$

The beta function comes up as a normalizing constant for beta distributions (Section 4.4.2). It's often useful to recognize the following identity:

$$B(\alpha_1, \alpha_2) = \int_0^1 x^{\alpha_1-1} (1-x)^{\alpha_2-1} dx$$



## B.2 The Poisson distribution

The POISSON DISTRIBUTION is a generalization of the binomial distribution in which the number of trials  $n$  grows arbitrarily large while the mean number of successes  $\pi n$  is held constant. It is traditional to write the mean number of successes as  $\lambda$ ; the Poisson probability density function is

$$P(y; \lambda) = e^{-\lambda} \frac{\lambda^y}{y!} \quad (y = 0, 1, \dots) \quad (\text{B.1})$$

The Gamma distribution is conjugate for the Poisson parameter  $\lambda$ , hence it is common to use a Gamma prior on  $\lambda$  in Bayesian inference.

## B.3 The hypergeometric distribution

One way of thinking of the binomial distribution is as  $n$  repeated draws from a bag with  $M$  marbles,  $\pi M$  of which are black and the rest of which are white; each outcome is recorded and the drawn marble is replaced in the bag, and at the end the total number of black marbles is the outcome  $k$ . This picture is often called SAMPLING WITH REPLACEMENT. The HYPERGEOMETRIC distribution is similar to this conception of the binomial distribution except that the marbles are not replaced after drawn—this is SAMPLING WITHOUT REPLACEMENT. The hypergeometric distribution has three parameters: the number of marbles  $M$ , the number of black marbles  $m$ , and the number of draws  $n$ ; the probability mass function on the number of “successes”  $X$  (black marbles drawn) is

$$P(X = r) = \frac{\binom{m}{r} \binom{M-m}{n-r}}{\binom{M}{n}}$$

In this book, the hypergeometric distribution comes up in discussion of Fisher’s exact test (Section 5.4.3).

## B.4 The chi-square distribution

Suppose that we have a standard normal random variable  $Z$ —that is,  $Z \sim N(0, 1)$ . The distribution that the quantity  $Z^2$  follows is called the CHI-SQUARE DISTRIBUTION with ONE DEGREE OF FREEDOM. This distribution is typically denoted as  $\chi_1^2$ .

If we have  $k$  independent random variables  $U_1, \dots, U_k$  such that each  $U_i \sim \chi_1^2$ , then the distribution of  $U = U_1 + \dots + U_k$  is the chi-squared with  $k$  degrees of freedom. This is denoted as  $U \sim \chi_k^2$ . The expectation of  $U$  is  $k$  and its variance is  $2k$ .

Figure B.1 illustrates the probability density functions for  $\chi^2$  distributions with various degrees of freedom. The  $\chi_1^2$  distribution grows asymptotically as  $x$  approaches 0, and  $\chi_2^2$

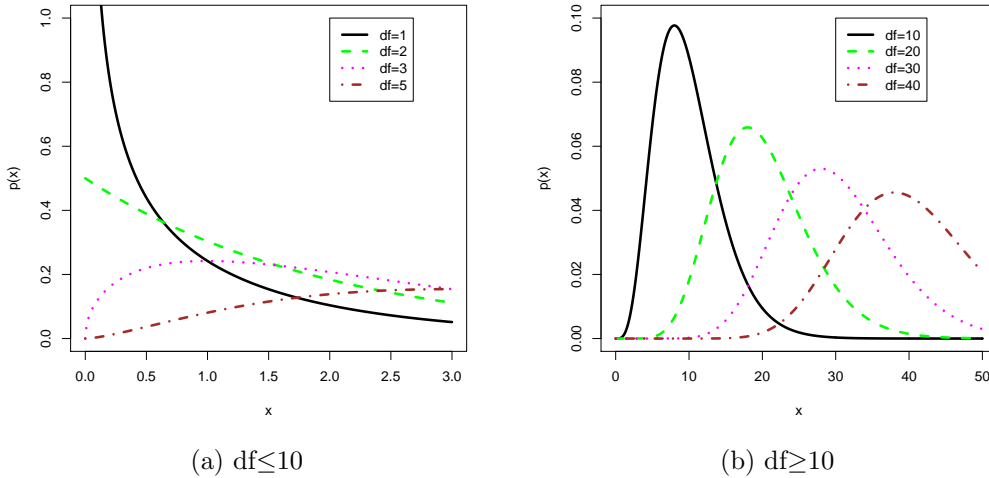


Figure B.1: The  $\chi^2$  distribution with various degrees of freedom

decreases monotonically, but all other  $\chi_k^2$  distributions have a mode for some positive  $x < k$ . As  $k$  grows large, more and more of the probability mass becomes located relatively close to  $x = k$ .

The key place where  $\chi^2$  variables arise is as the distribution of variance of a normal distribution. If we sample  $n$  points from  $\mathcal{N}(\mu, \sigma^2)$  (once again: that's a normal distribution with mean  $\mu$  and variance  $\sigma^2$ ), then the quantity

$$\frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \hat{\mu})^2$$

is distributed as  $\chi_{n-1}^2$ .

If  $U$  is distributed as  $\chi_k^2$ , then the distribution of the quantity  $1/U$  is called the INVERSE CHI-SQUARE DISTRIBUTION with  $k$  degrees of freedom. The inverse chi-square distribution is used in Bayesian inference as a conjugate prior (Section 4.4.3) for the variance of the normal distribution.

## B.5 The $t$ -distribution

Suppose once again that we have a standard normal random variable  $Z \sim N(0, 1)$ , and also that we have a chi-squared random variable  $U$  with  $k$  degrees of freedom. The distribution of the quantity

$$\frac{Z}{\sqrt{U/k}} \tag{B.2}$$

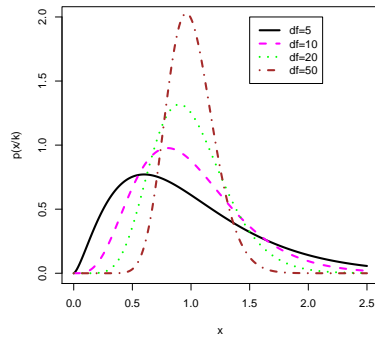


Figure B.2: The  $\chi^2$  distribution, normalized by degrees of freedom

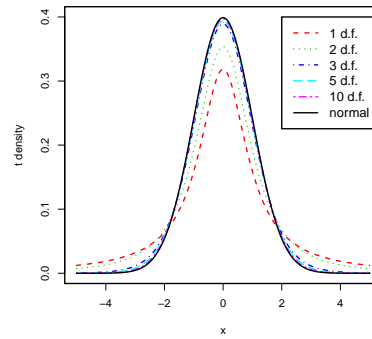


Figure B.3: The  $t$  distribution

is called the  $t$ -DISTRIBUTION WITH  $k$  DEGREES OF FREEDOM. It has expectation 0, and as long as  $k > 2$  its variance is  $\frac{k}{k-2}$  (it has infinite variance if  $k \leq 2$ ).

Figure B.3 shows the probability density functions for  $t$  distributions with varying degrees of freedom, together with the standard normal distribution for reference. The  $t$  distribution is *heavier-tailed* than the normal distribution, but even with 10 degrees of freedom the  $t$  distribution is already very close to the standard normal. As the degrees of freedom grow, the  $t$  distribution converges to the standard normal; intuitively, this is because  $\chi_k^2$  becomes more and more centered around  $k$ , so the quantity  $U/k$  in Equation B.2 converges to 1.

## B.6 The $F$ distribution

The  $F$  distribution is just the distribution of the normalized ratio of two independent normalized  $\chi^2$  random variables. More formally, if  $U \sim \chi_{k_1}^2$  and  $V \sim \chi_{k_2}^2$ , we have

$$F_{k_1, k_2} \sim \frac{U/k_1}{V/k_2} \quad (\text{B.3})$$

Here are a few things to note about the  $F$  distribution:

- The  $F$  distribution comes up mainly in frequentist hypothesis testing for linear models (Section 6.5).
- As  $k_1$  and  $k_2$  grow, all the probability mass in the  $F$  distribution converges to  $x = 1$ . Because the variance of a sample is distributed as a  $\chi^2$  random variable, the ratio of variances in linear models (as in Figure 6.9) can be compared to the  $F$  distribution.
- Consider the case where  $k_1 = 1$ . Since  $U$  is then the square of a standard normal random variable, a random variable with distribution  $F_{1, k_2}$  has the same distribution as the square of a random variable with distribution  $t_{k_2}$  (compare Equation (B.2)).

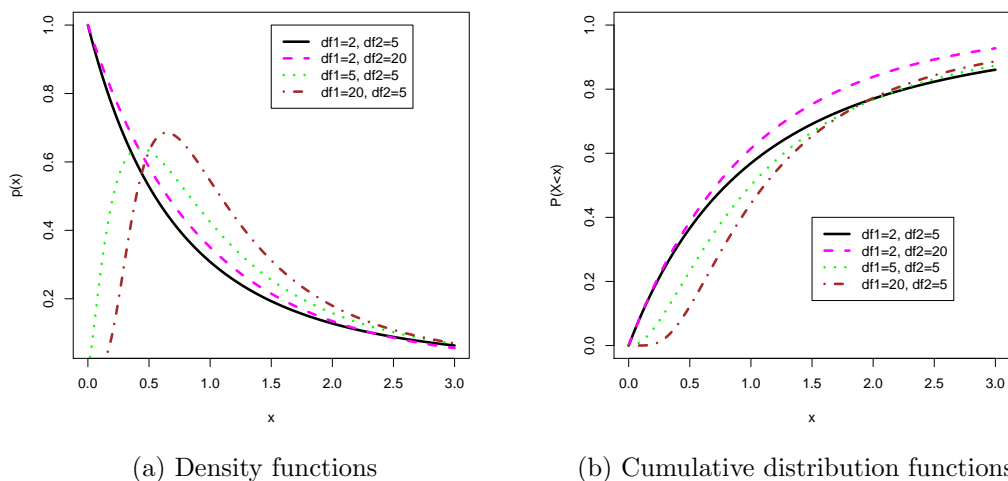


Figure B.4: Density and cumulative distribution functions for the  $F$  distribution

It is useful to play a bit with the  $F$  distribution to see what it looks like. Figure B.4 gives sample density and cumulative distribution functions for several choices of the degrees of freedom. In general, the cumulative distribution is more interesting and pertinent than the probability density function (unless you have an anomalously low  $F$  statistic).

## B.7 The Wishart distribution

Recall that the  $\chi^2$  distribution is used to place probability distributions over the inverse variance of a normal distribution (or of a sample from a normally-distributed population). The WISHART DISTRIBUTION is a multi-dimensional generalization of the  $\chi^2$  distribution, which generates inverse covariance matrices of multivariate normal distributions. Suppose that we have  $k$  independent observations from an  $n$ -dimensional multivariate normal distribution that itself has mean zero and covariance matrix  $\Sigma$ . Each observation  $\mathbf{z}_i$  can be written as  $\langle z_{i1}, \dots, z_{in} \rangle$ . If we write the matrix

$$\mathbf{Z} = \begin{bmatrix} z_{11} & z_{12} & \dots & z_{1n} \\ z_{21} & z_{22} & \dots & z_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ z_{k1} & z_{k2} & \dots & z_{kn} \end{bmatrix}$$

then the matrix  $\mathbf{Z}^T \mathbf{Z}$  follows a Wishart distribution with  $k$  degrees of freedom and scale matrix  $\Sigma$ .

If  $\mathbf{Z}$  is Wishart-distributed, then its inverse  $\mathbf{Z}^{-1}$  is said to be INVERSE WISHART-DISTRIBUTED. The inverse Wishart distribution is used in Bayesian inference as the conjugate prior (Section 4.4.3) for the covariance matrix of a multivariate normal distribution.

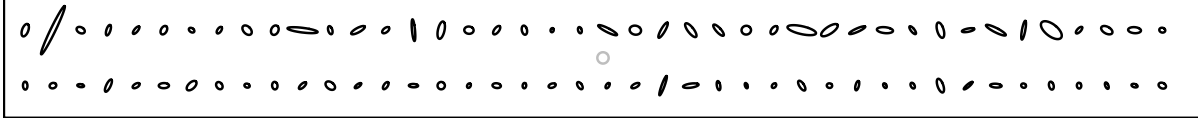


Figure B.5: Covariance-matrix samples from the two-dimensional inverse Wishart distribution with  $\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$  and  $k = 2$  (top row) or  $k = 5$  (bottom row), represented by their characteristic ellipses. The unit circle appears in gray in the center of the figure for reference.

Figure B.5 illustrates the inverse Wishart distribution for different degrees of freedom. Note that the variations in the covariance structure are more extreme when there are fewer degrees of freedom.

## B.8 The Dirichlet distribution

The DIRICHLET DISTRIBUTION is a generalization of the beta distribution (Section 4.4.2). Beta distributions are probability distributions over the success parameter  $\pi$  of a binomial distribution; the binomial distribution has two possible outcome classes. Dirichlet distributions are probability distributions over the parameters  $\pi_1, \dots, \pi_k$  of a  $k$ -class multinomial distribution (Section 3.4.1; recall that  $\pi_k$  is not a true model parameter as it is fully determined by  $\pi_1, \dots, \pi_{k-1}$ ). The Dirichlet distribution is characterized by parameters  $\alpha_1, \dots, \alpha_k$ , and  $\mathcal{D}(\pi_1, \dots, \pi_k)$  is defined as

$$\mathcal{D}(\pi_1, \dots, \pi_k) \stackrel{\text{def}}{=} \frac{1}{Z} \pi_1^{\alpha_1-1} \pi_2^{\alpha_2-1} \dots \pi_k^{\alpha_k-1}$$

where the normalizing constant  $Z$  is

$$Z = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)\dots\Gamma(\alpha_k)}{\Gamma(\alpha_1 + \alpha_2 + \dots + \alpha_k)}$$

By comparing with the beta function and beta distribution as defined in Sections 4.4.2 and B.1, it will be apparent that the beta distribution is a Dirichlet distribution in which  $k = 2$ . Just as there is a beta-binomial distribution giving the probability of obtaining  $y$  successes out of  $N$  draws from a binomial distribution drawn from a beta distribution, there is a Dirichlet-multinomial distribution that gives the probability of obtaining  $y_1, \dots, y_k$  outcomes in each of the  $k$  response classes respectively when taking  $N$  draws from a multinomial drawn from a Dirichlet distribution with parameters  $\alpha_1, \dots, \alpha_k$ . If we define  $\alpha = \sum_{i=1}^k \alpha_i$ , then the predictive distribution is (Leonard, 1977):

$$P(y_1, \dots, y_k) = \int_{\pi} P(y_1, \dots, y_k | \pi) P(\pi | \alpha_{1\dots k}) d\pi = \frac{\prod_{i=1}^k \binom{\alpha_i + y_i - 1}{\alpha_i}}{\binom{\alpha + N - 1}{\alpha}}$$

The special case of this predictive distribution is when we draw a multinomial distribution from the Dirichlet, and then draw one sample  $X$  from that multinomial distribution. The probability that that sample  $X$  has outcome class  $i$  is given by the value

$$P(X = i | \alpha_{1..k}) = \frac{\alpha_i}{\alpha}$$

This is often convenient for using Gibbs sampling to draw samples from the posterior distribution in Bayesian models which use Dirichlet priors over multinomial distributions. An example of this usage is given in Section 9.2.1.

The Dirichlet distribution has the following useful property. For any  $k$ -class Dirichlet distribution with parameters  $\alpha_1, \dots, \alpha_k$ , suppose we partition the  $k$  outcome classes into a smaller, new set of  $k' < k$  classes, with the  $j$ -th new class consisting of outcome classes  $c_{j1}, \dots, c_{jM_j}$ . The resulting distribution over the new set of  $k'$  outcome classes is *also* Dirichlet-distributed, with parameters  $\alpha_j = \sum_{i=1}^{M_j} \alpha_{ij}$ . [see also Dirichlet process in Section XXX; and give example here?]

# Appendix C

## The language of directed acyclic graphical models

Beginning with Chapter 8, this book makes considerable use of the formalism of DIRECTED ACYCLIC GRAPHICAL MODELS, or BAYESIAN NETWORKS (Bayes nets). In a few pages we cannot do justice to the diversity of work within this formalism, but this appendix introduces the formalism and a few critical accompanying concepts.

In the general case, graphical models are a set of formalisms for compactly expressing different types of conditional independence relationships between an ENSEMBLE of random variables. A graphical model on an ensemble  $X_1, \dots, X_n$  is literally a graph with one node for each random variable  $X_i$ , and in which each node may or may not be connected to each other node. The class of *directed* graphical models is those graphical models in which all the inter-node connections have a direction, indicated visually by an arrowhead. The class of directed *acyclic* graphical models, or DAGs (or Bayes nets), is those directed graphical models with no cycles—that is, one can never start at a node  $X_i$  and, by traverse edges in the direction of the arrows, get back to  $X_i$ . DAGs are the only type of graphical model that you’ll see in this book. Figure C.1 shows examples of several different types of graphical models.

### C.1 Directed graphical models and their interpretation

The structure of a given DAG encodes what conditional independencies hold among the variables in the ensemble  $X_1, \dots, X_n$ . First a bit of nomenclature. The PARENTS of a node  $X_i$  are the nodes that are pointing directly to it—in Figure C.1d, for example, the parents of  $X_5$  are  $X_3$  and  $X_4$ . The ANCESTORS of a node are all the nodes that can be reached from the node by traveling “upstream” on edges in the direction opposite to the arrows—in Figure C.1d, for example, all other nodes are ancestors of  $X_5$ , but  $X_4$  has no ancestors.

The set of connections between nodes in a DAG has a formal semantic interpretation whose simplest statement is as follows:

Any node  $X_i$  is conditionally independent of its non-descendants given its parents.

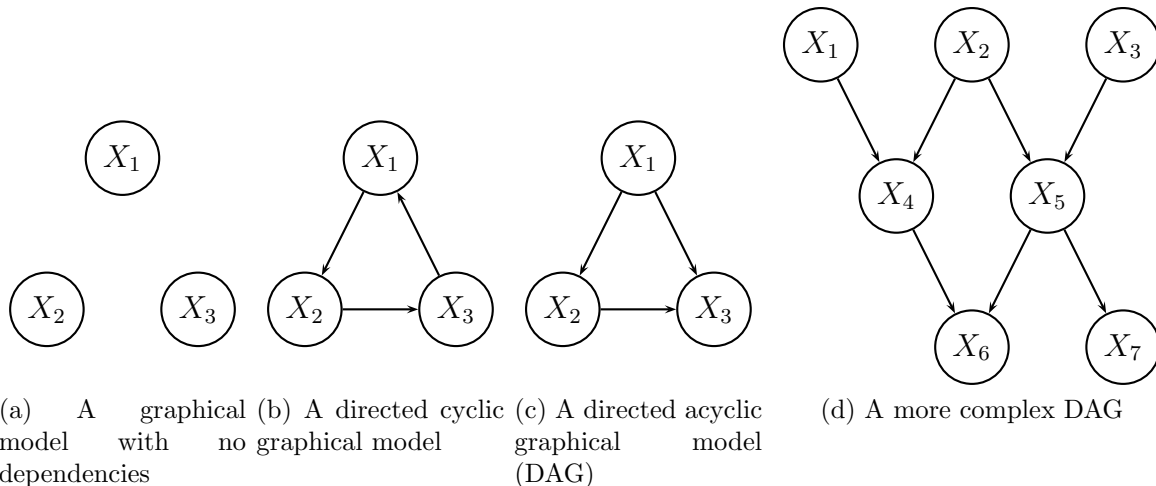


Figure C.1: Different classes of graphical models

In Figure C.1d, for instance, we have that:

$$X_6 \perp \{X_1, X_2, X_3, X_7\} \mid \{X_4, X_5\}$$

An important proviso is that—recalling from Chapter 2—conditional independencies can disappear with the accrual of new knowledge. In particular, two nodes are *not* conditionally independent of one another given a common descendent. So in Figure C.1d, for example,  $X_4$  has many conditional independencies given only its parents:

$$X_4 \perp \{X_3, X_5, X_7\} \mid \{X_1, X_2\}$$

but two of them go away when its child  $X_6$  is also given:

$$\begin{aligned} X_4 \not\perp X_3 \mid \{X_1, X_2, X_6\} \\ X_4 \not\perp X_5 \mid \{X_1, X_2, X_6\} \\ X_4 \perp X_7 \mid \{X_1, X_2, X_6\} \end{aligned}$$

A more complete statement of conditional independence in DAGs is given in Section C.2.

This statement of conditional independence simplifies the factorization of the joint probability distribution into smaller components. For example, we could simply use the chain rule (Section 2.4) to write the joint probability distribution for Figure C.1d as follows:

$$P(X_{1\dots 7}) = P(X_7|X_{1\dots 6})P(X_6|X_{1\dots 5})P(X_5|X_{1\dots 4})P(X_4|X_{1\dots 3})P(X_3|X_1, X_2)P(X_2|X_1)P(X_1)$$



but we can use the following conditional independencies, which can be read off the connectivity in the graph, to simplify this:

$$\begin{aligned}
P(X_7|X_{1\dots6}) &= P(X_7|X_5) \\
P(X_6|X_{1\dots5}) &= P(X_6|X_4, X_5) \\
P(X_5|X_{1\dots4}) &= P(X_5|X_3) \\
P(X_4|X_{1\dots3}) &= P(X_4|X_1, X_2) \\
P(X_3|X_1, X_2) &= P(X_3) \\
P(X_2|X_1) &= P(X_2)
\end{aligned} \tag{C.1}$$

giving us the following expression for the joint probability distribution

$$P(X_{1\dots7}) = P(X_7|X_5)P(X_6|X_5, X_4)P(X_5|X_3)P(X_4|X_1, X_2)P(X_3)P(X_2)P(X_1)$$

which is much simpler. These minimal conditional probability distributions seen in (C.1) are the components whose form needs to be specified in order to give a complete probabilistic model of a given domain.

[say something about proper indexing of variables?]

When conducting statistical inference in DAGs, it is often the case that we observe the more “downstream” variables and need to infer some of the more “upstream” variables. The catch is that the conditional probability distributions in the DAG are specified in terms of downstream variables given upstream variables. Conducting inference upstream, then, requires Bayesian inference (the reason that DAGs are often called “Bayes nets”). As an example, in Figure C.1d suppose that we observe (or choose via prior knowledge) all variables except  $X_4$ . To draw inferences about  $X_4$ , we’d use Bayes rule, targeting the downstream variable  $X_6$  for Bayesian inversion:

$$P(X_4|X_1, X_2, X_3, X_5, X_6, X_7) = \frac{P(X_6|X_{1\dots5}, X_7)P(X_4|X_{1\dots3}, X_5, X_7)}{P(X_6|X_{1\dots3}, X_5, X_7)}$$

We can now apply the conditional independencies in the graph to simplify all the numerator of the right-hand side:

$$= \frac{P(X_6|X_4, X_5)P(X_4|X_1, X_2)}{P(X_6|X_{1\dots3}, X_5, X_7)}$$

If we wanted to compute the denominator of Equation C.2, we’d need to do it by marginalizing over all possible values  $x_4$  that can be taken by  $X_4$ :

$$= \frac{P(X_6|X_4, X_5)P(X_4|X_1, X_2)}{\sum_{x_4} P(X_6|X_{1\dots5}, X_7)P(X_4|X_{1\dots3}, X_5, X_7)}$$

Applying the conditional independencies of the graph to the explicit marginalization reveals that  $X_3$  and  $X_7$  can be ignored:

$$= \frac{P(X_6|X_4, X_5)P(X_4|X_1, X_2)}{\sum_{x_4} P(X_6|X_4, X_5)P(X_4|X_1, X_2)}$$

If we now drop the explicit marginalization, we obtain the simplest characterization of Bayesian inference on  $X_4$  available for this graph:

$$P(X_4|X_1, X_2, X_3, X_5, X_6, X_7) = \frac{P(X_6|X_4, X_5)P(X_4|X_1, X_2)}{P(X_6|X_1, X_2, X_5)} \quad (\text{C.2})$$

## C.2 Conditional independence in DAGS: d-separation†

We have already briefly described the intuitive picture for when conditional independence holds in a DAG: given its parents, a node is conditionally independent of all of its non-descendants. However, we also saw that such conditional independencies can be broken when more information is conditioned on. In this section, we give the comprehensive criterion by which conditional independence can be assessed in any DAG. This criterion is known as D-SEPARATION (Pearl, 1988, Section 3.3).

Consider two disjoint subsets  $A$  and  $B$  of nodes in a DAG. A PATH between  $A$  and  $B$  is simply a sequence of edges that, when taken together, connects some node in  $A$  with some node in  $B$  (note that this definition doesn't require that the arrows along the path all point in the same direction). Any node on a given path is said to have CONVERGING ARROWS if two edges on the path connect to it and point to it. A node on the path is said to have NON-CONVERGING ARROWS if two edges on the path connect to it, but at least one does not point to it. (Note that the starting and ending nodes on the path are each connected to by only one edge on the path, so are not said to have either converging or non-converging arrows.)

Now consider a third subset  $C$  of nodes in the DAG, disjoint from both  $A$  and  $B$ .  $C$  is said to d-separate  $A$  and  $B$  if for every path between  $A$  and  $B$ , one of the following two properties holds:

1. there is some node on the path with converging arrows which *is not* in  $C$ ; or
2. there is some node on the path whose arrows do not converge and which *is* in  $C$ .

If  $C$  d-separates  $A$  and  $B$ , then  $A$  and  $B$  must be conditionally independent given  $C$ . If  $C$  does not d-separate  $A$  and  $B$ , then  $A$  and  $B$  are not in general conditionally independent.

Figure C.2 illustrates the canonical cases of d-separation and of failure of d-separation. In Figures C.2a, we have d-separation:  $C$  is on the path between  $A$  and  $B$ , and it does not have converging arrows. Therefore if  $C$  is known, then  $A$  and  $B$  become conditionally independent:

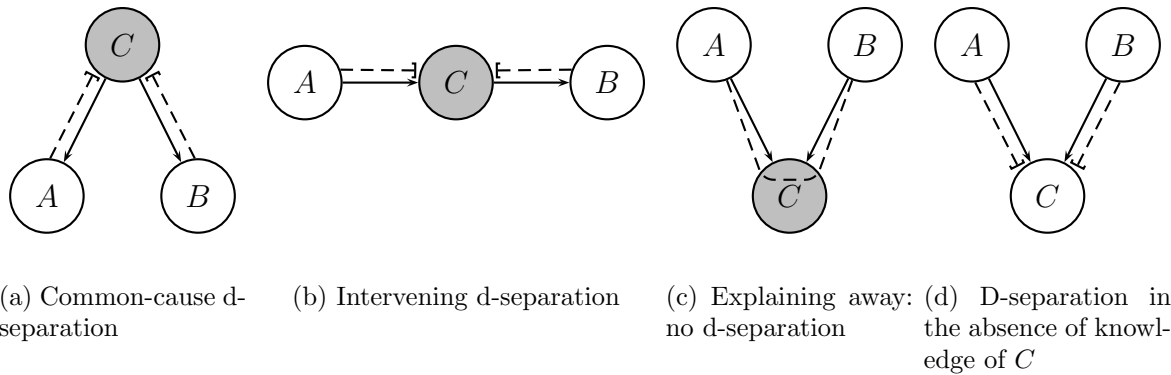


Figure C.2: Examples of d-separation and of failure of d-separation.  $C$  blocks the path between  $A$  and  $B$  in Figures C.2a, C.2b, and C.2d, but not in Figure C.2c.

$A \perp B \mid C$ .<sup>1</sup> This configuration is sometimes called “common cause” d-separation: if  $A$  and  $B$  are the outcomes of two tosses of a possibly unfair coin, then knowing the coin’s weighting ( $C$ ) renders the tosses independent.

The same holds of Figure C.2b:  $C$  is on the path between  $A$  and  $B$ , and doesn’t have converging arrows, so  $A \perp B \mid C$ . This configuration is often known as “indirect cause”: if I know my mother’s genome ( $C$ ), then the respective contents of my genome ( $B$ ) and my mother’s mother’s genome ( $A$ ) become conditionally independent.

In Figures C.2c and C.2d, on the other hand,  $C$  is on the path between  $A$  and  $B$  but it has converging arrows. Therefore  $C$  does not d-separate  $A$  and  $B$ , so  $A \not\perp B \mid C$  (Figure C.2c). This configuration is often known as “common effect”: a signal ( $C$ ) indicating whether the tosses of two fair coins ( $A$  and  $B$ ) came up on the same side renders the two tosses conditionally dependent. However, *not* having seen this signal leaves the two tosses independent. In the language of graphical models, d-separation, and conditional independence, we have  $A \perp B \mid \emptyset$  (Figure C.2d).

For further reading, some key sources are Pearl (1988, 2000); Jordan (1998); Russell and Norvig (2003, Chapter 14); Bishop (2006, Chapter 8).

<sup>1</sup>Technically, since d-separation is a property holding among *sets* of nodes, we should write  $\{A\} \perp \{B\} \mid \{C\}$ ; but for simplicity we drop the braces as a slight abuse of notation when a set consists of exactly one node.

# Appendix D

## Dummy chapter

# Bibliography

- Abney, S., McAllester, D., and Pereira, F. (1999). Relating probabilistic grammars and automata. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 542–549.
- Agresti, A. (2002). *Categorical Data Analysis*. Wiley, second edition.
- Agresti, A. (2007). *An Introduction to Categorical Data Analysis*. Wiley, second edition.
- Alexopolou, T. and Keller, F. (2007). Locality, cyclicity, and resumption: At the interface between the grammar and the human sentence processor. *Language*, 83(1):110–160.
- Aslin, R. N., Saffran, J. R., and Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, 9:321–324.
- Baayen, H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge.
- Baayen, R. H. (2001). *Word Frequency Distributions*. Springer.
- Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412.
- Baayen, R. H., Piepenbrock, R., and Gilkerson, L. (1995). The CELEX lexical database (2nd edition). CD-ROM.
- Bailey, C.-J. (1973). *Variation and Linguistic Theory*. Washington: Center for Applied Linguistics.
- Bailey, T. M. and Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods. *Journal of Memory and Language*, 44:568–591.
- Balota, D. A. and Spieler, D. H. (1998). The utility of item level analyses in model evaluation: A reply to seidenberg and plaut. *psysci*, 9:238–240.
- Bates, D. (2007). Linear mixed model implementation in `lme4`. Manuscript, University of Wisconsin, 15 May 2007.

- Benor, S. B. and Levy, R. (2006). The chicken or the egg? A probabilistic analysis of English binomials. *Language*, 82(2):233–278.
- Beyer, K. (1986). *The Aramaic language, its distributions and subdivisions*. Vandenhoeck & Ruprecht.
- Bicknell, K., Elman, J. L., Hare, M., McRae, K., and Kutas, M. (2010). Effects of event knowledge in processing verbal arguments. *jml*, 63:489–505.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
- Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.
- Bloom, P. (2001). *How Children Learn the Meanings of Words*. Academic Press: San Diego.
- Bod, R. (1992). A computational model of language performance: Data oriented parsing. In *Proceedings of COLING*.
- Bod, R. (1998). *Beyond Grammar: An Experience-Based Theory of Language*. CSLI Press.
- Bod, R. (2006). An all-subtrees approach to unsupervised parsing. In *Proceedings of ACL-COLING*.
- Bolinger, D. (1962). Binomials and pitch accent. *Lingua*, 11:34–44.
- Bresnan, J., Cueni, A., Nikitina, T., and Baayen, H. (2007). Predicting the dative alternation. In Boume, G., Kraemer, I., and Zwarts, J., editors, *Cognitive Foundations of Interpretation*, pages 69–95. Amsterdam: Royal Netherlands Academy of Science.
- Cavall-Sforza, L. and Feldman, M. (1981). *Cultural Transmission and Evolution: a Quantitative Approach*. Princeton University Press.
- Cedergren, H. J. and Sankoff, D. (1974). Variable rules: Performance as a statistical reflection of competence. *Language*, 50(2):333–355.
- Chen, S. and Goodman, J. (1998). An empirical study of smoothing techniques for language modeling. Technical report, Computer Science Group, Harvard University.
- Chi, Z. (1999). Statistical properties of probabilistic context-free grammars. *Computational Linguistics*, 25(1):131–160.
- Cho, T. and Keating, P. (2009). Effects of initial position versus prominence in english. *jphon*, 37:466–485.
- Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge.
- Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 12:335–359.

- Cohen, J., Cohen, P., West, S. G., and Aiken, L. S. (2003). *Applied Multiple Regression/correlation Analysis for the Behavioral Sciences*. Lawrence Earlbaum, third edition.
- Collins, P. (1995). The indirect object construction in english: an informational approach. *Linguistics*, 33:35–49.
- Cooper, W. and Ross, J. (1975). World order. In *Proceedings of the Chicago Linguistic Society*, pages 63–111.
- Cox, R. T. (1946). Probability, frequency and reasonable expectation. *American Journal of Physics*, 14(1):1–13.
- Cramer, H. (1946). *Mathematical Methods of Statistics*. Princeton, N.J.: Princeton University Press.
- Cuetos, F. and Mitchell, D. C. (1988). Cross-linguistic differences in parsing: Restrictions on the use of the Late Closure strategy in Spanish. *Cognition*, 30(1):73–105.
- Daumé, III, H. and Marcu, D. (2006). Domain adaptation for statistical classifiers. *jair*, 26:101–126.
- Dubey, A., Keller, F., and Sturt, P. (2008). A probabilistic corpus-based model of syntactic parallelism. *Cognition*, 109(3):326–344.
- Gahl, S. (2008). “time” and “thyme” are not homophones: Lemma frequency and word durations in a corpus of spontaneous speech. Under review.
- Gazdar, G., Klein, E., Pullum, G., and Sag, I. (1985). *Generalized Phrase Structure Grammar*. Harvard.
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC, second edition.
- Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge.
- Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. In *Proceedings of the IEEE International Conference on Acoustics, Speech, & Signal Processing*, pages 517–520.
- Goldwater, S., Griffiths, T. L., and Johnson, M. (2007). Distributional cues to word segmentation: Context is important. In *Proceedings of the 31st Boston University Conference on Language Development*.
- Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101:5228–5235.

- Griffiths, T. L., Steyvers, M., and Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, 114(2):211–244.
- Hale, K. (1983). Warlpiri and the grammar of non-configurational languages. *Natural Language and Linguistic Theory*, 1(1):5–47.
- Harrell, Jr, F. E. (2001). *Regression Modeling Strategies*. Springer.
- Hawkins, J. A. (1994). *A Performance Theory of Order and Constituency*. Cambridge.
- Hayes, B. and Wilson, C. (2007). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, 39:379–440.
- Healy, M. (2000). *Matrices for Statistics*. Oxford, second edition.
- Jeffrey, R. (2004). *Subjective Probability: The Real Thing*. Cambridge University Press.
- Jordan, M. I., editor (1998). *Learning in Graphical Models*. Cambridge, MA: MIT Press.
- Kempen, G. and Harbusch, K. (2004). How flexible is constituent order in the midfield of german subordinate clauses? a corpus study revealing unexpected rigidity. In *Proceedings of the Conference on Linguistic Evidence*.
- Kroch, A. (1989). Reflexes of grammar in patterns of language change. *Language Variation and Change*, 1:199–244.
- Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., and Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science*, 9(2):F13–F21.
- Landauer, T. K. (1986). How much do people remember? some estimates of the quantity of learned information in long-term memory. *cogsci*, 10:477–493.
- Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato’s problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240.
- Leonard, T. (1977). A Bayesian approach to some multinomial estimation problems. *Journal of the American Statistical Association*, 72(360):869–874.
- Levy, R. (2002). The statistical distribution of English coordinate noun phrases: Parallelism and weight effects. Presented at NWAV 31.
- Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106:1126–1177.
- Lieberman, A. M., Delattre, P., and Cooper, F. (1958). Some cues for the distinction between voiced and voiceless stops in initial position. *Language and Speech*, 1:153–167.



- Lisker, L. and Abramson, A. S. (1967). Some effects of context on voice onset time in English. *Language and Speech*, 10(1):1–28.
- Liu, J. S. (1994). The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, 89(427):958–966.
- Maindonald, J. and Braun, J. (2007). *Data Analysis and Graphics using R*. Cambridge, second edition.
- Malkiel, Y. (1959). Studies in irreversible binomials. *Lingua*, 8:113–160.
- McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall, second edition.
- Müller, G. (1997). Beschränkungen für binomialbildungen im deutschen. *Zeitschrift für Sprachwissenschaft*, 16(1):25–51.
- Nagy, W. E. and Anderson, R. C. (1984). How many words are there in printed school English? *Reading Research Quarterly*, 19(3):304–330.
- Nosofsky, R. (1986). Attention, similarity and the identification-categorization relationship. *Journal of Experimental Psychology: Human Perception & Performance*, 115:39–57.
- O’Shannessy, C. (2006). *Language contact and children’s bilingual acquisition: learning a mixed language and Warlpiri in northern Australia*. PhD thesis, University of Sydney.
- Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 2 edition.
- Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge.
- Peterson, G. E. and Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 24(2):175–184.
- Peterson, P. G. (1986). Establishing verb agreement with disjunctively conjoined subjects: Strategies vs principles. *Australian Journal of Linguistics*, 6(2):231–249.
- Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer.
- Pinker, S. and Birdsong, D. (1979). Speakers’ sensitivity to rules of frozen word order. *Journal of Verbal Learning and Verbal Behavior*, 18:497–508.
- Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer, second edition.
- Roland, D., Dick, F., and Elman, J. L. (2007). Frequency of basic English grammatical structures: A corpus analysis. *Journal of Memory and Language*, 57:348–379.

- Ross, J. R. (1967). *Constraints on Variables in Syntax*. PhD thesis, MIT.
- Russell, S. and Norvig, P. (2003). *Artificial Intelligence: a Modern Approach*. Prentice Hall, second edition.
- Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996a). Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928.
- Saffran, J. R., Newport, E. L., and Aslin, R. N. (1996b). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, 35:606–621.
- Sag, I. A., Gazdar, G., Wasow, T., and Weisler, S. (1985). Coordination and how to distinguish categories. *Natural Language and Linguistic Theory*, 3:117–171.
- Sankoff, D. and Labov, W. (1979). On the uses of variable rules. *Language in Society*, 8:189–222.
- Scha, R. (1990). Language theory and language technology: Competence and performance. In *Computertoepassingen in de Neerlandistiek*. Almere: Landelijke Vereniging van Neerlandici.
- Scholes, R. J. (1966). *Phonotactic Grammaticality*. Mouton.
- Smith, N. A. and Johnson, M. (2007). Weighted and probabilistic context-free grammars are equally expressive. *Computational Linguistics*, 33(4):477–491.
- Spieler, D. H. and Balota, D. A. (1997). Bringing computational models of word naming down to the item level. *psychsci*, 6:411–416.
- Stanford, J. (2008). A sociotonec analysis of Sui dialect contact. *Language Variation and Change*, 20(3):409–450.
- Stram, D. O. and Lee, J. W. (1994). Variance components testing in the longitudinal mixed-effects model. *Biometrics*, 50:1171–1177.
- Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, 50:86–132.
- Tool, M. (1949). *Resurrection Road*. St. Martin’s Minotaur.
- Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, fourth edition.
- Wang, W. S.-Y. and Minett, J. (2005). The invasion of language: emergence, change and death. *Trends in Ecology and Evolution*, 20(5):263–269.
- Wasow, T. (2002). *Postverbal Behavior*. CSLI.

- Weide, R. L. (1998). The CMU pronouncing dictionary. Version 0.6; available online at <http://www.speech.cs.cmu.edu/cgi-bin/cmudict>.
- Weiner, E. J. and Labov, W. (1983). Constraints on the agentless passive. *Journal of Linguistics*, 19:29–58.
- Werker, J. F. and Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, 7:49–63.
- Zaenen, A., Carletta, J., Garretson, G., Bresnan, J., Koontz-Garboden, A., Nikitina, T., O'Connor, M. C., and Wasow, T. (2004). Animacy encoding in English: Why and how. In *Proceedings of the ACL 2004 Workshop on Discourse Annotation*.