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 π 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.[1]

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

[1] 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 {*interest, principal*}) then mathematically our linguist is saying that $P(\texttt{p}) = P(\texttt{i}) = \frac{1}{2}$.

2.2 Sample Spaces

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

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

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

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

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 Ω is a function from events in Ω 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:

(1)

	Pronoun	**Not Pronoun**
Object **Preverbal**	0.224	0.655
Object **Postverbal**	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 explicity 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:

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

The chain rule

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

$$P(E_1 \cap E_2 \cap \cdots \cap E_n) = P(E_n|E_1 \cap E_2 \cap \cdots \cap E_{n-1}) \ldots P(E_2|E_1)P(E_1) \qquad (2.3)$$

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:

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

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

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

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:

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

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.

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

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 $A_1, A_2, \ldots, A_n$, 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

$$P(A \cap B|C) = P(A|C)P(B|C) \tag{2.12}$$

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 Ω of a probability space to a finite, or countably infinite, set of real numbers ($\mathbb{R}$).[2] 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.

The relationship between the sample space Ω, a probability space P on Ω, and a discrete random variable X on Ω 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 $\Omega = \{\texttt{pp}, \texttt{pi}, \texttt{ip}, \texttt{ii}\}$. 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 ω 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 $P(X)$:

ω	$X(\omega)$	$P(X)$
`pp`	2	$\frac{1}{4}$
`pi`	1	$\frac{1}{2}$
`ip`	1	
`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 Ω mapped to y are in that equivalence class. Intuitively, a random variable can be thought of as focusing

[2] A set S is COUNTABLY INFINITE if a one-to-one mapping exists between the integers $0, 1, 2, \ldots$ 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 {*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 π. π is the probability of achieving success on a single Bernoulli trial, and can take any value between 0 and 1 (inclusive); π 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 $c_1, \ldots, c_r$. This is a MULTINOMIAL TRIAL, and just as the Bernoulli distribution has 1 parameter, the distribution for multinomial trials has $r - 1$ parameters $\pi_1, \ldots, \pi_{r-1}$ determining the probability that a trial will fall into each of the classes:

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

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 π_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 $\pi_{\texttt{e}} = 0.126$, $\pi_{\texttt{t}} = 0.099$, $\pi_{\texttt{a}} = 0.082$, 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 π has the following (very simple!) cumulative distribution function:

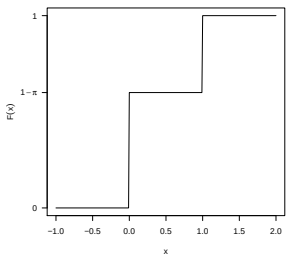


Figure 2.1: The cumulative distribution function for a Bernoulli random variable with parameter π

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

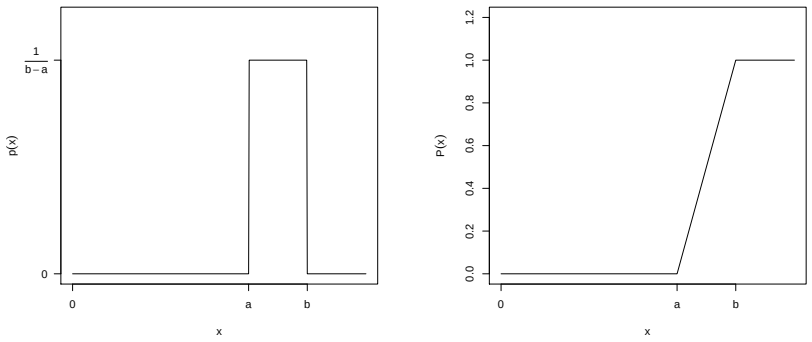
which is illustrated in Figure 2.1.

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

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

$p(x) = \frac{1}{b-a}$ 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:

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

2.7.2 Change of variables for continuous probability densities

Typically, the range of a continuous random variables is some kind of metric space used to quantify events in the real world. In many cases, there may be more than one possible metric of interest. In phonetics, for example, pitch is sometimes expressed directly in units of frequency, namely Hertz (cycles per second), but sometimes measured in log-Hertz instead. The justifications for log-Hertz measurement include that in music relative changes in pitch are constant in frequency *ratio*, so that adding a constant value c to a log-Hertz measurement yields the same change in musical pitch regardless of starting frequency; and that in many cases vowels tend to be constant in the ratios among their first three formats (e.g., Lloyd (1890); Peterson (1961); Miller (1989); Hillenbrand et al. (1995)). [3] If one wants to convert a probability density from one unit of measurement to another, one needs to make a CHANGE OF VARIABLES.

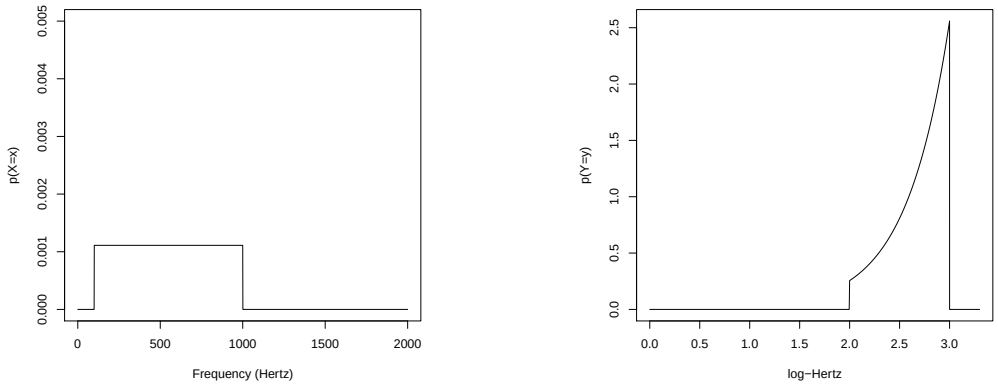
In general, if some random variable X has the probability density p and the random variable Y is defined such that $X = g(Y)$, then the probability density of Y is

$$p(Y = y) = p(X = g(y)) \frac{dg}{dy}(g(y))$$

or, more succinctly, For example, suppose that one has a random variable X with a uniform distribution over the Hertz frequency range [100,1000]. To convert this to a distribution over log-frequencies, let us call the new random variable Y defined that $Y = log_{10}X$, so that the log-Hertz range of Y is [2,3]. This means that $X = g(y) = 10^Y$ and so $\frac{dg}{dy} = 10^Y \log 10$ (see Section A.5). Figures **??** and **??** illustrate this probability density in the two units of measurement.

This example illustrates an important point: that the pdf of a continuous random variable does *not* need to be bounded above by 1; changing variables from Hertz to log-Hertz led to

[3] A formant is a peak of energy in the acoustic frequency spectrum of a vowel production.



a density as high as 2.5. Since the range in which the density exceeds 1 is small (in the log-Hertz scale), this is not a problem as the total probability mass still integrates to 1. Note that technically, since probability mass is unitless but continuous probability densities are over variables that have units, probability densities have units—in the case of the uniform distribution in the example above, the unit is "per cycle per second". And in fact, for *any* continuous probability function there will always be some change of variables based on a new unit of measurement that will lead to density exceeding 1 somewhere on the range of the new random variable! This is in contrast to the probability mass function of a discrete random variable, which is unitless must be bounded above by 1.

2.7.3 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.

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 $F(x)$".

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, γ_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 γ_2 will indicating the preference for S to precede V, and a third parameter γ_3 indicating the preference for O to precede V. For simplicity, we'll let each γ_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 γ_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:

$$(1) S \prec O, V \prec S, O \prec V \qquad\qquad (2) O \prec S, S \prec V, V \prec O$$

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. In Table 2.1 we make totally explicit the eight logically possible combinations of

	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	$\prec$	$\succ$	$\prec$	**impossible**	$\gamma_1(1-\gamma_2)\gamma_3$
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	$\succ$	$\prec$	$\succ$	**impossible**	$(1-\gamma_1)\gamma_2(1-\gamma_3)$
7	$\succ$	$\succ$	$\prec$	OVS	$(1-\gamma_1)(1-\gamma_2)\gamma_3$
8	$\succ$	$\succ$	$\succ$	VOS	$(1-\gamma_1)(1-\gamma_2)(1-\gamma_3)$

Table 2.1: The unnormalized distribution for {S,O,V} constituent-order model

Order	SOV	SVO	VSO	OSV	OVS	VOS
# Languages	566	488	95	25	11	4
Relative frequencies	0.476	0.410	0.080	0.021	0.009	0.003
Probabilities in constituent-order model	0.414	0.414	0.103	0.046	0.011	0.011

Table 2.2: Empirical frequencies of dominant constituent ordering among {S,O,V} for 1189 languages, taken from the World Atlas of Language Structures (Dryer, 2011; languages reported as lacking dominant order omitted). Model probabilities are for $\gamma_1 = 0.9, \gamma_2 = 0.8, \gamma_3 = 0.5$).

three pairwise precedences, the two that are contradictory, and the values assigned by F to each of the eight.

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

This is an interesting model: because it has only three parameters, it is not as expressive as an arbitrary multinomial distribution over six classes (we would need five parameters for that; see Section 2.5.2). It's an empirical question whether it's good for modeling the kind of word-order frequencies across the languages of the world. The first two rows of Table 2.2 show the empirical frequencies of the six logically possible orderings of subject, object, and verb; the two subject-initial orderings are by far the most common, with VSO a distant third and the other orders all quite rare. If we wanted to produce a probability distribution that looked like empirical frequencies, intuitively we might set γ_1 close to 1, since S nearly always precedes O; γ_2 close to 1 as well, since S nearly always precedes V, but lower than γ_1, since the former generalization is stronger than the second; and γ_3 around $\frac{1}{2}$, since V precedes O about as often as it follows it. The third row of Table 2.2 shows the probabilities in the constituent-order model obtained with such a setting, of $\gamma_1 = 0.9, \gamma_2 = 0.8, \gamma_3 = 0.5$. It is a pretty good qualitative fit: it fails to differentiate SOV from SVO probability but reproduces the overall shape of the empirical relative frequency distribution reasonably well. In fact, this three-parameter constituent-order model can achieve even better fits to word-order-

frequency data than we see in Table 2.2; the principles according to which the optimal fit can be determined will be introduced in Chapter 4, and the model is revisited in Exercise 4.6.

Problem 2.6 in the end of this chapter revisits the question of whether the parameters γ_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.

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 μ, 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 π, for example, the possible outcomes are 0 and 1, so we have

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

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 2.8).

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:

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

or equivalently

$$\text{Var}(X) = E[X^2] - E[X]^2 \tag{2.20}$$

(see Exercise 3.1). The variance is often denoted σ^2 and its positive square root, σ, is known as the STANDARD DEVIATION.

If you *rescale* a random variable by defining $Y = a + bX$, then $\text{Var}(Y) = b^2X$. This is part of what is known as LINEARITY OF THE EXPECTATION, which will be introduced in full in Section 3.3.1.

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 (2.20) and summing over the possible outcomes as in Equation (2.17) (recall that the expectation for a Bernoulli random variable is π):

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

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 2.9).

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 μ and the variance σ^2. (Sometimes the STANDARD DEVIATION σ is used instead of the variance to parameterize the normal distribution.) Its probability density function is:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{2.21}$$

This expression seems intimidating at first glance, but it will become familiar with time.[4] It can be divided into three components:

[4] $\exp[x]$ 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.

- $\frac{1}{\sqrt{2\pi\sigma^2}}$ 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.

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 2.7.3 in expressing the answer in terms of the cumulative distribution function:

$$P(225\text{Hz} \leq \text{F1} \leq 267\text{Hz}) = \int_{225}^{267} p(x)\,\mathrm{d}x \tag{2.22}$$

$$= \int_{-\infty}^{267} p(x)\,\mathrm{d}x - \int_{-\infty}^{225} p(x)\,\mathrm{d}x \tag{2.23}$$

$$= F(267) - F(225) \tag{2.24}$$

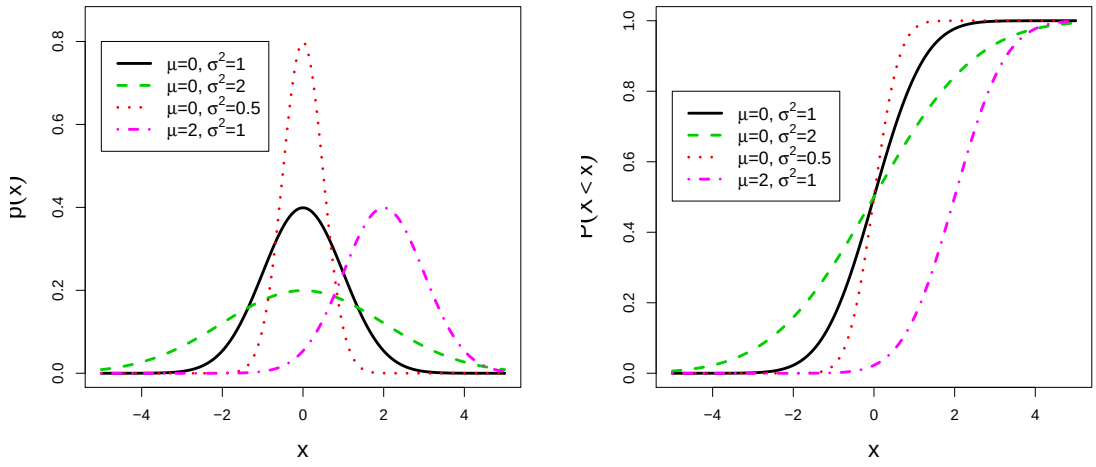
With the use of standard statistical software we can find the values of the cumulative distributions function at 267 and 225, which gives us our answer:

$$= 0.5 - 0.12 = 0.38 \tag{2.25}$$

(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 $p(x) = \frac{1}{\sqrt{2\pi}} e^{\left[-x^2/2\right]}$.



(a) Probability density function

(b) Cumulative distribution function

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

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.

2.11.1 Discrete probability densities: relative-frequency estimation

Suppose that we are interested in estimating the Bernoulli parameter π 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 $\hat{\pi} = \frac{7}{10}$.[5] 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

[5]The $\hat{}$ symbol above the π indicates that this is an *estimate* of the parameter π which may or not be the true underlying parameter value.

	⟨SB, DO⟩	⟨SB, IO⟩	⟨DO, SB⟩	⟨DO, IO⟩	⟨IO, SB⟩	⟨IO, DO⟩	Total
Count	478	59	1	3	20	9	570
Relative Freq.	0.839	0.104	0.001	0.005	0.035	0.016	

Table 2.3: 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 direct object.

of different values. Table 2.3, 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 2.3). Once again we simply divide the count of each outcome by the total count, as shown in the bottom line of Table 2.3. 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.

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 ɑ by adult male speakers in the classic study of Peterson and Barney (1952).[6] 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 weakness. 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

[6]The measurements reported in this dataset are rounded off to the nearest Hertz, so in Figure 2.4a they are jittered to break ties.

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.[7] Hence we have

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

in the discrete case, or

$$\int_x K(x_i, x)\, dx = 1 \tag{2.27}$$

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 $x_1, \ldots, x_n$, 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:

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} K(x_i, x) \tag{2.28}$$

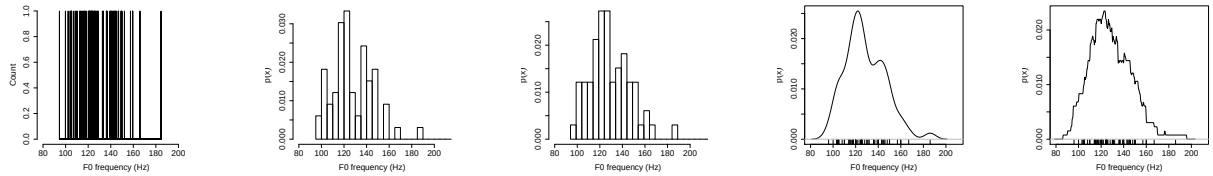
It is up to the researcher to choose the particular kernel function. Here, we give an example for a continuous random variable; in the next section we give an example for discrete kernel density estimation.

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 σ. The standard deviation is sometimes called the BANDWIDTH of the kernel and denoted as b. The kernel density estimate from a set of observations $x_1, \ldots, x_n$ 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.

[7] 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.



(a) Raw observations (b) 5Hz-bin histogram starting at 95Hz (c) 5Hz-bin histogram starting at 94Hz (d) Normal-kernel density estimate, $b = 5$ (e) Rectangular-kernel density estimate, $b = 20$

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

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.[8]

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 that can help determine how good a given choice of bandwidth may be. Exercise 2.11 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.

[8] 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.

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/,/æ/,/ʌ/, and/ʊ/. 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, $D(/\text{pæt}/, /\text{tʊt}/) = 2$. Less similar phoneme sequences should have a lower score in our kernel, so let us define our kernel as

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

We have used proportionality rather than equality here because of the requirement (Equation (2.26)) 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:

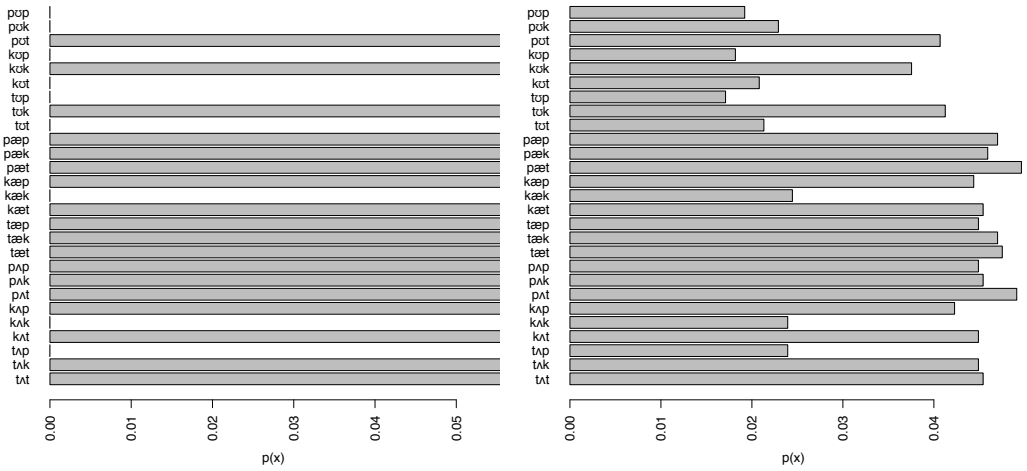
$$K(x_i, x') = \frac{1}{Z_i} \frac{1}{(1 + D(x, x'))^3}$$

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 (2.28), computing the probability of an outcome using kernel density estimation (KDE) involves iterating explicitly over the entire set of observations $\boldsymbol{y}$. 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



(a) Relative frequency estimation (b) Kernel density estimate

Figure 2.5: Kernel density estimation for lexical forms

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 $\boldsymbol{y}$ 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 $\hat{y}_1, \ldots, \hat{y}_n$. 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 {*night, 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	`n`	0.45
2	40	35	`d`	0.53
3	48	27	`d`	0.36
4	38	37	`d`	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 $\{night, day\}$ the best classification decision is d if $\hat{P}(\texttt{d}) > 0.5$; but $\hat{P}(\texttt{d}) = 0.51$ and $\hat{P}(\texttt{d}) = 0.99$ 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 $\boldsymbol{y}$ is simply

$$P(\boldsymbol{y}|\hat{P}) \tag{2.29}$$

The likelihood is sometimes viewed as a function of a set of observations $\boldsymbol{y}$, 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

$$P(\boldsymbol{y}|\hat{P}) = \prod_{i=1}^{n} P(y_i|\hat{P}) \tag{2.30}$$

Since likelihoods can be very small, and datasets can be very large, explicitly computing the product in Equation (2.30) 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 (2.31) below:

$$\log P(\boldsymbol{y}|\hat{P}) = \sum_{i=1}^{n} \log P(y_i|\hat{P}) \tag{2.31}$$

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 $b \approx 10$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

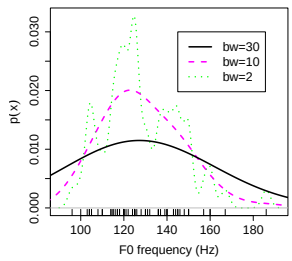


Figure 2.6: F0 formant kernel density estimates for normal kernels of different bandwidths.

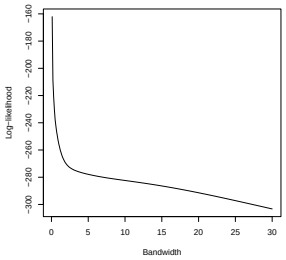


Figure 2.7: Overfitting when evaluating on training set: F0 formant log-likelihood increases unboundedly.

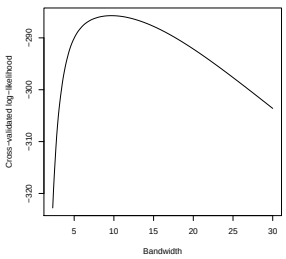
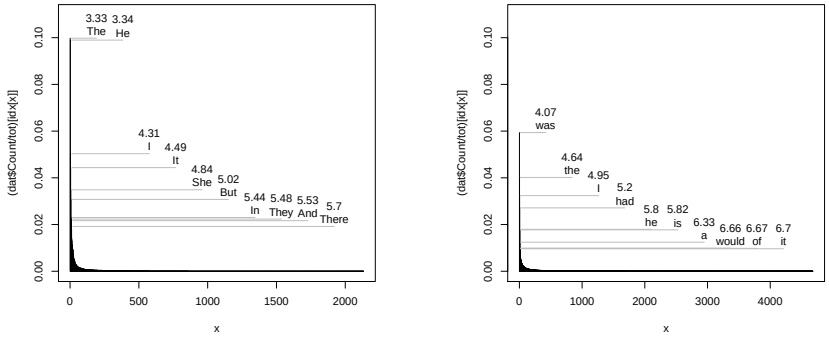


Figure 2.8: Cross-validated log-likelihood reveals the optimal bandwidth.



(a) First words

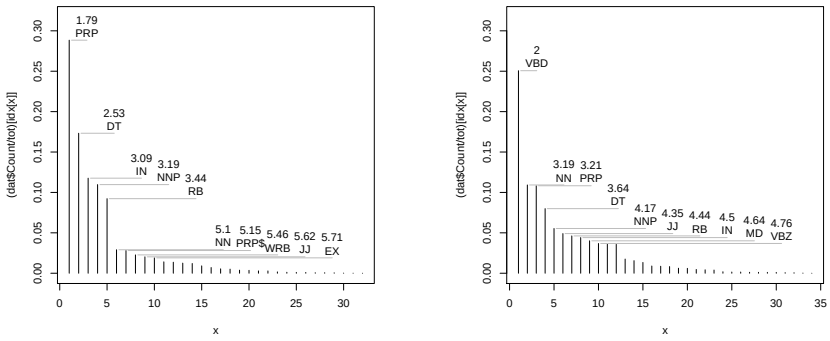
(b) Second words

Figure 2.9: The relative frequency distributions of the first and second words of sentences of the parsed Brown corpus

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 Surprisal, entropy, and relative entropy

One of the most fundamental views of probability is as quantifying our degree of uncertainty about events whose outcomes are unknown. Intuitively, the more uncertain we are as to an event's outcome, the more broadly distributed will be the probability mass over the event's possible outcomes. Likewise, we can say that learning the event's outcome gives us *information* about the world that we did not previously have. The quantity of information



(a) First-word POS (b) Second-word POS

Figure 2.10: The relative frequency distributions of the parts of speech of the first and second words of sentences of the parsed Brown corpus

conveyed by the outcome x of a discrete random variable is often measured by the SURPRISAL, alternatively called the SHANNON INFORMATION CONTENT or SELF-INFORMATION, of the outcome, defined as $\frac{1}{\log_2 P(x)}$ or equivalently $-\log_2 P(x)$; the unit of surprisal is the BIT, which is the amount of information conveyed by the flip of a fair coin. The surprisal of an event is the minimum number of bits required to convey the event's occurrence given knowledge of the underlying probability distribution from which the event was generated.

As an example, Figure **??** shows the relative frequencies (in descending order) of all words observed as the first word of at least one sentence in the parsed Brown corpus, excluding punctuation (Marcus et al., 1994). The ten most common of these words are labeled in the graph. If we oversimplify slightly and take these relative frequencies to be the true underlying word probabilities, we can compute the surprisal value that each word would have if it is seen as the first word in a new sentence drawn at random from the same corpus. These surprisal values appear above each of the ten labeled words. We can see that although there are thousands of different words attested to start sentences in this corpus, none of the ten most common conveys more than six bits of information (for reference, $2^6 = 64$).

The expected surprisal of a discrete random variable X, or the average information that an outcome of X conveys, is known as its ENTROPY $H(X)$, defined as:

$$H(X) = \sum_x P(x) \log_2 \frac{1}{P(x)} \tag{2.32}$$

or equivalently

$$= -\sum_x P(x) \log_2 P(x) \tag{2.33}$$

Applying this definition we find that the entropy of the distribution over first words in the sentence is 7.21, considerably higher than the entropy for second words, which is 9.51. The

precise values of these figures should be taken with a considerable grain of salt, because the data are quite sparse (most words only occur a handful of times, and many possible words don't appear at all), but they suggest that there is considerably more uncertainty about the second word in a sentence than about the first word of the sentence (recall, of course, that these probabilities do not at this point take into account any information about the context in which the word appears other than how many words preceded it in the sentence). Figure 2.10 shows relative frequency plots for the parts of speech of these first and second words (note, interestingly, that while PRP, or preposition, is the most frequent part of speech for first words, the most frequent word is a determiner). Repeating the calculation for parts of speech yields entropies of 3.33 and 3.8 for the first and second words respectively. Once again the entropy for the second word is greater than the first word, though by a smaller amount than in the word-level calculation (perhaps due to the fact that second words are likelier than first words to be open-class parts of speech, which are much easier to predict at the part-of-speech level than at the word-specific level).

Finally, consider the case where one has two different probability distributions P and Q over the same event space. Note that in the definition of entropy in Equation (2.32) the same probability function is used twice; one could imagine carrying out a similar calculation using each of P and Q once:

$$\sum_x P(x) \log_2 \frac{1}{Q(x)} \tag{2.34}$$

$$\tag{2.35}$$

This is known as CROSS ENTROPY. It is useful to think of the distribution Q appearing inside the logarithm as a *guess* distribution and the other distribution P as the *true*, or *reference*, distribution. Cross entropy quantifies how many bits are required on average to convey an event drawn from P when one does not know P and one's best guess of the distribution is Q. To determine how much worse this is than using the true distribution (and it is never better!), we can subtract out the entropy of P; this gives us what is called the RELATIVE ENTROPY or KULLBACK-LEIBLER DIVERGENCE (or KL divergence) from Q to P:

$$D(P||Q) = \sum_x P(x) \log_2 \frac{P(x)}{Q(x)} \tag{2.36}$$

The KL divergence can be thought of as the penalty, in bits, incurred by coding outcomes from P using Q. It is never negative and is zero *only* when $P = Q$. In our most recent example, if we call the two distributions over part-of-speech tags P_1 for first words and P_2 for second words, we find that the KL divergence $D(P_1||P_2) = 0.92$. (there are some parts of speech appearing in the second-word position which do not appear in the first-word position, such as the possessive clitic *'s*, which is represented as a separate word in the Penn Treebank, so that we cannot take the KL divergence from P_1 to P_2; it would be infinite.) This KL divergence is well over a third the size of the entropy of the true distributions, indicating that the part of speech distributions are very different for first and second words.

Surprisal, entropy, and relative entropy are essential conceptual building blocks in INFORMATION THEORY (Shannon, 1948), which among many other important ideas and results includes the SOURCE CODING THEOREM, which gives theoretical bounds on the compressibility of any information source, and the NOISY CHANNEL THEOREM, which gives theoretical bounds on the possible rate of error-free information transfer in noisy communication systems. Cover and Thomas (1991) is an authoritative text on many key areas of information theory; MacKay (2003) is another accessible text covering these two theorems.

```
> roundN <- function(x,decimals=2,fore=5) sprintf(paste("%",fore,".",decimals,"f",sep=
```

2.13 Exercises

Exercise 2.1: Conditional independence and set intersection†

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**

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`? What about `tee`? [Hint: see Section 2.5.2; perhaps peek at Section A.8 as well.]

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 γ_i analogously to the success parameter of a single Bernoulli trial, so that the probability that $S \prec O$ is γ_1, the probability that $S \prec V$ is γ_2, and the probability that $O \prec V$ is γ_3. Given the mass function F actually used in the example, is the probability that $S \prec O$ actually γ_1? Show your work.

Exercise 2.7: More on the constituent-order model

Play around with specific parameter values for the constituent-order model to get a feel for it. We know that it is more constrained than a general six-class multinomial distribution, since it has only three parameters instead of five. Qualitatively speaking, what kinds of distributions over the six logically possible word orders is it incapable of modeling?

Exercise 2.8: Expectation of a uniform random variable*

Prove mathematically that the expectation of a uniform random variable X on $[a, b]$ is $E(X) = \frac{a+b}{2}$. (This involves a simple integration; consult A.6 if you need a refresher.)

Exercise 2.9: Variance of a uniform random variable**

Prove that the variance of a continuous, uniformly distributed random variable X on $[a, b]$ is $\frac{(b-a)^2}{12}$. [Sections 2.7.1 and 2.9.2]

Exercise 2.10: Normal distributions♣

For adult female native speakers of American English, the distribution of first-formant frequencies for the vowel [ɛ] 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 [ɛ] for a randomly selected adult female native speaker of American English will be between 555Hz and 697Hz?

Exercise 2.11: Assessing optimal kernel bandwidth through cross-validation†,⌨

Use leave-one-out cross-validation to calculate the cross-validated likelihood of kernel density estimates (using a normal kernel) of adult male speaker [ɑ] 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 [ɑ] and [i] similar to each other? Explain your results.

Exercise 2.12: Kernel density estimation and change of variables

Complete Exercise 2.11, but this time change the formant frequency measurements from Hertz to log-Hertz before carrying out bandwidth selection. Once you're done, compare the optimal bandwidths and the corresponding density estimates obtained using log-Hertz

measurements with that obtained using Hertz measurements. How different are they?

Exercise 2.13: Kernels in discrete linguistic spaces[‡]

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 2.14: Exemplar-based model of phonotactic knowledge[‡]

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 estimtes 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.