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. [1] 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 $X_1, X_2, \cdots, X_N$, the joint density is written as

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

or simply

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

for short.

[1] 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 $F(x_1, \cdots, x_n)$, is defined as the probability of the set of random variables all falling at or below the specified values of X_i:[2]

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

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:

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

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[3] over the possible outcomes of the random variables that are not of interest. For two variables X, Y:

[2] Technically, the definition of the multivariate cumulative distribution function is

$$F(x_1, \cdots, x_n) \overset{\text{def}}{=} P(X_1 \leq x_1, \cdots, X_N \leq x_n) = \sum_{\vec{x} \leq \langle x_1, \cdots, x_N \rangle} p(\vec{x}) \qquad \text{[Discrete]} \tag{3.3}$$

$$F(x_1, \cdots, x_n) \overset{\text{def}}{=} P(X_1 \leq x_1, \cdots, X_N \leq x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_N} p(\vec{x}) dx_N \cdots dx_1 \qquad \text{[Continuous]} \tag{3.4}$$

[3] or integral in the continuous case

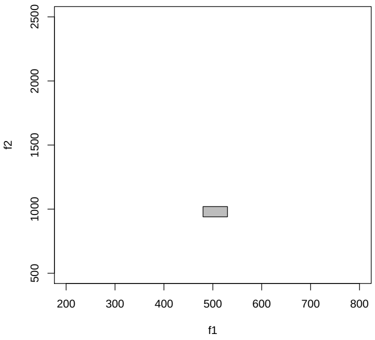


Figure 3.1: The probability of the formants of a vowel landing in the grey rectangle can be calculated using the joint cumulative distribution function.

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

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:

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

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 f sums of random variables

3.3.1 Linearity of the expectation

Linearity of the expectation is an extremely important property and can 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 introduc 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 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:

(1)

		Coding for Y		
		0	1	
Coding for X		**Pronoun**	**Not Pronoun**	
0	Object **Preverbal**	0.224	0.655	.879
1	Object **Postverbal**	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 3.2 asks you to prove this).

3.3.3 Covariance and scaling random variables

What happens to $Cov(X, Y)$ when you scale X? Let $Z = a + bX$. It turns out that the covariance with Y increases by b (Exercise 3.4 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 3.3).

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 commmon to use the CORRELATION ρ as a standardized form of covariance:

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

```
[1]  0.020653248 -0.018862690 -0.009377172  0.022384614
```

In the word order & pronominality example above, where we found that the covariance of verb-object word order and object pronominality was 0.01, we can re-express this relationship as a correlation. We recall that the variance of a Bernoulli random variable with success parameter π is $\pi(1-\pi)$, so that verb-object word order has variance 0.11 and object pronominality has variance 0.18. The correlation between the two random variables is thus $\frac{0.01}{\sqrt{0.11\times0.18}} = 0.11$.

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 $Z = X_1 + \cdots + X_n$. Then

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

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 π, 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 π.

For example, it is intuitively obvious that the mean of a binomially distributed r.v. Y with parameters n and π is πn. 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 $k = 0, 1, \dots, n$: [4]

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

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 π, so we have:

$$E(Y) = \sum_{i}^{n} E(X_i) \tag{3.9}$$

$$= \sum_{i}^{n} \pi = \pi n \tag{3.10}$$

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 $n\pi(1 - \pi)$.

[4] 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 $n \times (n - 1) \times \cdots \times 1$ for $n = 1, 2, \dots$, and as 1 for $n = 0$.

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 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 π_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 $\pi_r \overset{\text{def}}{=} 1 - \sum_{i=1}^{r-1} \pi_i$, 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:

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

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

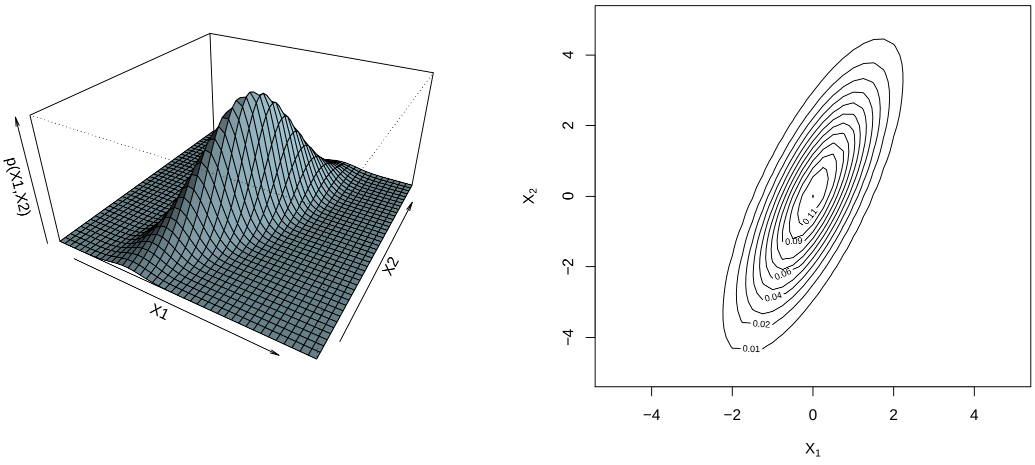
3.5 Multivariate normal distributions

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 μ and variance σ^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 μ of length N, and (2) a symmetric COVARIANCE MATRIX (or variance-covariance matrix) $\mathbf{\Sigma}$ 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 $\mathbf{\Sigma}$ are the variances of the individual X_i and must be non-negative. In this situation we sometimes say that $X_1, \dots, X_N$ are JOINTLY NORMALLY DISTRIBUTED.

The probability density function for the multivariate normal distribution is most easily expressed using matrix notation (Section A.9); the symbol $\mathbf{x}$ stands for the vector $\langle x_1, \dots, x_n \rangle$:

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

For example, a bivariate normal distribution ($N = 2$) over random variables X_1 and X_2 has two means μ_1, μ_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 Y_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 σ_{11}^2 and σ_{22}^2 are



(a) Perspective plot

(b) Contour plot

Figure 3.2: Visualizing the multivariate normal distribution

simply the variances of X_1 and X_2 respectively (the subscripts appear doubled for notational consistency). The term σ_{12}^2 is the covariance between the two axes. [5] Figure 3.2 visualizes a bivariate normal distribution with $\mu = (0, 0)$ and $\mathbf{\Sigma} = \begin{pmatrix} 1 & 1.5 \\ 1.5 & 4 \end{pmatrix}$. 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.

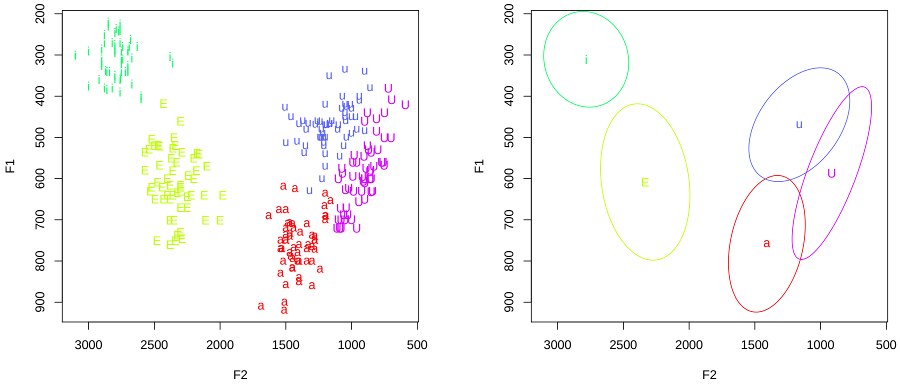
The multivariate normal distribution is very useful in modeling multivariate data such as the distribution of multiple formant frequencies in vowel production. As an example, Figure 3.3 shows how a large number of raw recordings of five vowels in American English can be summarized by five "characteristic 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.

[5] The probability density function works out to be

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

Note that if σ_{11} is much larger than σ_{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)

3.5.1 The sum of jointly normal random variables

Yet another attractive property of the multivariate normal distribution is that the sum of a set of jointly normal random variables is itself a normal random variable. The mean and variance of the sum can be computed based on the formulae given in Sections 3.3.1 and 3.3.5. So if $\langle X_1, \dots, X_n \rangle$ are jointly normal with mean $\langle \mu_1, \dots, \mu_n \rangle$ and covariance matrix Σ, then $Z = X_1 + \dots + X_n$ is normally distributed with mean $\sum_{i=1}^{n} \mu_i$ and variance $\sum_{i=1}^{n} \sigma_i^2 + \sum_{i \neq j} \sigma_{ij}$.

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 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 3.12 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 Joint entropy, conditional entropy, and mutual information

In Section 2.12 we introduced the basic information-theoretic ideas of surprisal and entropy. With multivariate probability, there is not much more to say about surprisal: all there is to say is that the surprisal of the joint outcome of multiple random variables is the log of the inverse of the joint probability of outcomes:

$$\log \frac{1}{P(x_1, x_2, \ldots, x_n)} \qquad \text{or} \qquad -\log P(x_1, x_2, \ldots, x_n). \tag{3.13}$$

However, there is much more to say about entropies. In the rest of this section we will limit the discussion to cases where there are two random variables X and Y, but most of what is discussed can be generated to collections of arbitrary quantities of random variables.

We begin by defining the JOINT ENTROPY of X and Y analogously from the surprisal of a joint outcome:

$$H(X,Y) = \sum_{x,y} P(x,y) \log \frac{1}{P(x,y)} \tag{3.14}$$

What gets really interesting is when we break down the joint entropy into its constituent parts. We start by imagining situations in which we obtain knowledge of X while remaining ignorant of Y. The average entropy that Y will have after we learn about X is called the CONDITIONAL ENTROPY of Y given X and is notated as follows:

$$H(Y|X) = \sum_{x} P(x) \sum_{y} P(y|x) \log_2 \frac{1}{P(y|x)} \tag{3.15}$$

where $P(x)$ is the marginal probability of x. Note that this equation follows simply from the definition of expectation. Recall that in Section 2.12 we showed the distributions and entropies of non-punctuation words and their corresponding parts of speech. Returning to this example and slightly modifying the dataset (now excluding all sentences in which either the first or the second word was a punctuation term, a more stringent criterion), we find that the entropy of the part of speech for the second word is 3.66 and that its conditional entropy given the first word's part of speech is 2.43. That is, the first word removes about a third of the entropy of the second word!

Next, we can ask how much information we would lose regarding the joint distribution of X and Y if we were to treat the two variables as independent. Recall once again from Section 2.12 that the KL divergence from Q to P measures the penalty incurred by using Q to approximate P. Here, let us define $Q(x,y) = P_X(x)P_Y(y)$ where P_X and P_Y are the marginal probabilities for X and Y respectively. The KL divergence from Q to P is known as the MUTUAL INFORMATION between X and Y and is defined as

$$I(X;Y) = \sum_{x,y} P(x,y) \log \frac{P(x,y)}{P_X(x)P_Y(y)} \tag{3.16}$$

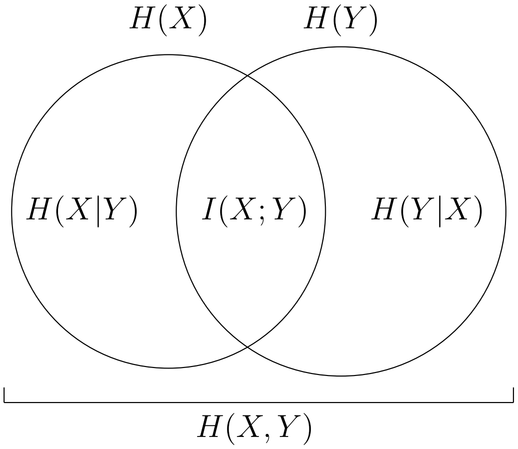


Figure 3.4: Entropy and mutual information for two random variables as a Venn diagram. Circle sizes and positions reflect the entropies of our example, where X is the first-word part of speech and Y is the second-word part of speech.

In our example, the mutual information between the parts of speech for the first and second words comes out to 1.23. You may notice that the three numbers we have just seen stand in a very simple relationship: $3.66 = 2.43 + 1.23$. This is no coincidence! In general, given any two random variables X and Y, the entropy of Y can *always* be decomposed as precisely the sum of the mutual information—which measures how much X tells you about Y—and the conditional entropy—which measures how much X *doesn't* tell you about Y:

$$H(Y) = I(X;Y) + H(Y|X) \qquad \text{and likewise} \qquad H(X) = I(X;Y) + H(X|Y). \tag{3.17}$$

There is one more remarkable decomposition to be had. In our example, the entropy of the first-word part of speech is 3.38, and the joint entropy for the two words is 5.81. In general, the joint entropy of X and Y is the sum of the individual variables' entropies minus the mutual information—which measures the *redundancy* between X and Y:

$$H(X,Y) = H(X) + H(Y) - I(X;Y) \tag{3.18}$$

In our case, 3.38 + 3.66 - 1.23 = 5.81. This decomposition arises from the original definition of mutual information as the coding penalty incurred for assuming independence between two variables.

In closing this section, let us notice that mutual information comes up in both an *asymmetric* decomposition—in the decomposition of $H(Y)$ as how much information X gives about Y—and in a *symmetric* decomposition—in the relationship between a joint entropy and the marginal entropies. For two random variables, the complete set of relations among the joint entropies, individual-variable entropies, conditional entropies, and mutual information can be depicted in a Venn diagram, as in Figure 3.4. The relations described in this section are well worth reviewing repeatedly, until they become second nature.

3.8 Exercises

Exercise 3.1: Simpler formula for variance.

Show from the definition of the variance as $\mathrm{Var}(X) \equiv E[(X - E(X))^2]$ that it can equivalently be written as $\mathrm{Var}(X) = E[X^2] - E[X]^2$, which we stated without proof in Section 2.9.2. [Section 3.3.1]

Exercise 3.2: 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 $\mathrm{Cov}(X, Y) = 0$ 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 3.3: ♣

- 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 $\mathrm{Var}(Z) = b^2\mathrm{Var}(X)$.

Exercise 3.4

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

Exercise 3.5

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

Exercise 3.6

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 π, 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

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

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

Exercise 3.7

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 3.8

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

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

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 3.9: 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 3.10

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 3.11

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 $(X + Y)$? [Section 3.3.5]

Exercise 3.12

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.

1