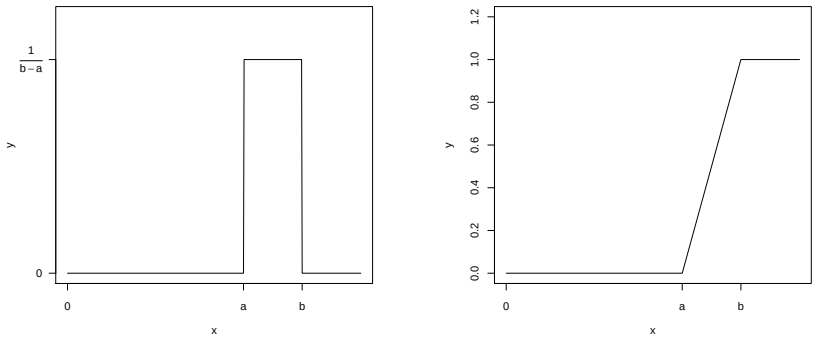

(a) Probability density function (b) Cumulative distribution function

Figure 2.3: The probability density function and cumulative distribution function of the uniform distribution with parameters a and b

2.8 Expected values and variance

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

2.8.1 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

$$E(X) = \sum_i x_i P(x_i) \tag{2.13}$$

For a Bernoulli random variable X_π with parameter π, for example, the possible outcomes are 0 and 1, so we have

$$E(X_\pi) = 0 \times (1 - \pi) + 1 \times \pi \tag{2.14}$$
$$= \pi \tag{2.15}$$

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

$$E(X) = \int_{-\infty}^{\infty} x \; p(x) dx \tag{2.16}$$

For example, a uniformly-distributed random variable $X_{(a,b)}$ with parameters a and b has expectation

$$E(X_{(a,b)}) = \int_{-\infty}^{a} x\ p(x)\,dx + \int_{a}^{b} x\ p(x)\,dx + \int_{b}^{\infty} x\ p(x)\,dx \tag{2.17}$$

$$= 0 + \int_{a}^{b} x\frac{1}{b-a}\,dx + 0 \tag{2.18}$$

$$= \frac{x^2}{2}\frac{1}{b-a}\Bigg|_{a}^{b} \tag{2.19}$$

$$= \frac{b^2-a^2}{2}\frac{1}{b-a} \tag{2.20}$$

$$= \frac{b+a}{2} \tag{2.21}$$

which corresponds nicely to the intuition that the expectation should be in the middle of the allowed interval.

2.8.2 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:

$$\mathrm{Var}(X) = E[(X - E(X))^2] \tag{2.22}$$

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

Variance of Bernoulli and uniform distributions

The Bernoulli distribution's variance needs to be calculated explicitly; recall that its expectation is π:

$$E[((X) - E(X))^2] = \sum_{x\in\{0,1\}} (x-\pi)^2\ P(x) \tag{2.23}$$

$$= \pi^2(1-\pi) + (1-\pi)^2 \times \pi \tag{2.24}$$

$$= \pi(1-\pi)[\pi + (1-\pi] \tag{2.25}$$

$$= \pi(1-\pi) \tag{2.26}$$

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 Homework XXX).

2.9 Joint probability distributions

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 ($\mathbb{R}$). 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 really quite simple: probability distributions can be generalized over multiple random variables at once, in which case they are called JOINT PROBABILITY DISTRIBUTIONS (jpd's). If a jpd is over N random variables at once then it maps from the sample space to $\mathbb{R}^N$, which is short-hand for real-valued VECTORS of dimension N. Notationally, for random variables $X_1, X_2, \cdots, X_N$, the joint probability density function is written as

$$p(X_1 = x_1, X_2 = x_2, \cdots, X_N = x_n)$$

or simply

$$p(x_1, x_2, \cdots, x_n)$$

for short.

Whereas for a single r.v., the cumulative distribution function is used to indicate the probability of the outcome falling on a segment of the real number line, the JOINT CUMULATIVE PROBABILITY DISTRIBUTION function indicates the probability of the outcome falling in a region of N-dimensional space. The joint cpd, 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:[3]

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

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 $480\text{Hz} \leq f_1 530\text{Hz}, 940\text{Hz} \leq f_2 \leq 1020\text{Hz}$ is given below:

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

[3] Technically, the definition of the multivariate cpd is then

$$F(x_1, \cdots, x_n) \stackrel{\text{def}}{=} P(X_1 \leq x_1, \cdots, X_N \leq x_n) = \sum_{\vec{x} \leq \langle x_1, \cdots, x_N \rangle} p(\vec{x}) \quad \text{[Discrete]} \quad (2.27)$$

$$F(x_1, \cdots, x_n) \stackrel{\text{def}}{=} P(X_1 \leq x_1, \cdots, X_N \leq x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_N} p(\vec{x}) dx_N \cdots dx_1 \quad \text{[Continuous]} \quad (2.28)$$

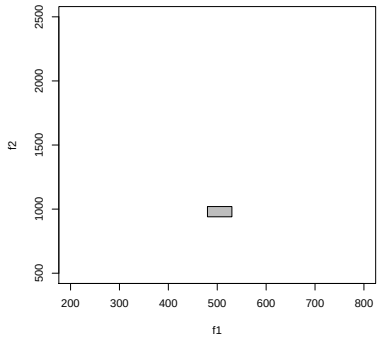

Figure 2.4: The probability of the formants of a vowel landing in the grey rectangle can be calculated using the joint cumulative distribution function.

and visualized in Figure 2.4 using the code below.

```
> plot(c(),c(),xlim=c(200,800),ylim=c(500,2500),xlab="f1",ylab="f2")
> rect(480,940,530,1020,col=8)
```

2.10 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[4] over the possible outcomes of the r.v.'s that are not of interest. For two variables X, Y:

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

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

2.11 Covariance

The COVARIANCE between two random variables X and Y is defined as follows:

[4] or integral in the continuous case

$$Cov(X,Y) = E[(X - E(X))(Y - E(Y))]$$

Simple example:

(2)

		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	

Each of X and Y can be treated as a Bernoulli random variable with arbitrary codings of 1 for **Postverbal** and **Not Pronoun**, and 0 for the others. As a resunt, we have $\mu_X = 0.121$, $\mu_Y = 0.762$. The covariance between the two is:

$$\begin{aligned}
&(0 - .121) \times (0 - .762) \times .224 && (0,0)\\
+&(1 - .121) \times (0 - .762) \times 0.014 && (1,0)\\
+&(0 - .121) \times (1 - .762) \times 0.0655 && (0,1)\\
+&(1 - .121) \times (1 - .762) \times 0.107 && (1,1)\\
=&0.0148
\end{aligned}$$

If X and Y are conditionally independent given our state of knowledge, then the covariance between the two is zero

In `R`, we can use the `cov()` function to get the covariance between two random variables, such as word length versus frequency across the English lexicon:

```
> cov(x$Length,x$Count)
[1] -42.44823
> cov(x$Length,log(x$Count))
[1] -0.9333345
```

The covariance in both cases is *negative*, indicating that longer words tend to be less frequent. If we shuffle one of the covariates around, it eliminates this covariance:

`order()` plus `runif()` give a nice way of randomizing a vector.

```
> cov(x$Length,log(x$Count)[order(runif(length(x$Count)))])
[1] 0.006211629
```

The covariance is essentially zero now.

Two important asides: the variance of a random variable X is just its covariance with itself:

$$\text{Var}(X) = \text{Cov}(X,X) \tag{2.29}$$

and any two random variables X and Y that are conditionally independent given our state of knowledge have covariance $\text{Cov}(X,Y) = 0$.

2.11.1 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:[5]

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

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

2.11.2 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:

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

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

2.12 Properties of Expectation and Variance

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.

[5]The reason for this is as follows. By linearity of expectation, $E(Z) = a + bE(X)$. This gives us

$$\begin{aligned}
Cov(Z, Y) &= E[(Z - a + bE(X))(Y - E(Y))] \\
&= E[(bX - bE(X))(Y - E(Y))] \\
&= E[b(X - E(X))(Y - E(Y))] \\
&= bE[(X - E(X))(Y - E(Y))] && \text{[by linearity of expectation]} \\
&= bCov(X, Y) && \text{[by linearity of expectation]}
\end{aligned}$$

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

$$E(Y) = a + \sum_i b_i E(X_i) \tag{2.30}$$

This is incredibly convenient. We'll demonstrate this convenience by introducing the binomial distribution in the next section.

Variance of the sum of random variables

What is the the variance of the sum of random variables $X_1 + \cdots + X_n$. We have

$$\begin{aligned}
\text{Var}(X_1 + \cdots + X_n) &= E\left[(X_1 + \cdots + X_n - E\left(X_1 + \cdots + X_n\right))^2\right] && (2.31)\\
&= E\left[(X_1 + \cdots + X_n - (\mu_1 + \cdots + \mu_n))^2\right] \text{ (Linearity of the expectation)} && (2.32)\\
&= E\left[((X_1 - \mu_1) + \cdots + (X_n - \mu_n))^2\right] && (2.33)\\
&= E\left[\sum_{i=1}^{n}(X_i - \mu_i)^2 + \sum_{i \neq j}(X_i - \mu_i)(X_j - \mu_j)\right] && (2.34)\\
&= \sum_{i=1}^{n} E\left[(X_i - \mu_i)^2\right] + \sum_{i \neq j} E\left[(X_i - \mu_i)(X_j - \mu_j)\right] \text{ (Linearity of the expectation)} && (2.35)\\
&= \sum_{i=1}^{n} \text{Var}(X_i) + \sum_{i \neq j} \text{Cov}(X_i, X_j) \text{(Definition of variance \& covariance)} && (2.36)
\end{aligned}$$

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.

2.13 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 $k = 0, 1, \ldots, n$:

$$P(Y = k) = \binom{n}{k} \pi^k (1 - \pi)^{n-k} \tag{2.37}$$

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:

$$E(Y) = \sum_i^n E(X_i) \tag{2.38}$$

$$= \sum_i^n \pi = \pi n \tag{2.39}$$

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

2.13.1 The multinomial distribution

The MULTINOMIAL DISTRIBUTION is the generalization of the binomial distribution to $r \geq 2$ possible outcomes. The r-class multinomial is a sequence of r random variables $X_1, \ldots, X_r$ whose joint distribution is characterized by r parameters: a size parameter n denoting the number of trials, and $r - 1$ parameters $\pi_1, \ldots, \pi_{r-1}$, 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 $\pi_r \stackrel{\text{def}}{=} 1 - \sum_{i=1}^{r-1} \pi_i$, but this is not a real parameter of the family because it's completely determined by the other parameters.) The (joint) probability mass function of the multinomial looks like this:

$$P(X_1 = n_1, \cdots, X_r = n_r) = \binom{n}{n_1 \cdots n_r} \prod_{i=1}^{r} \pi_i \tag{2.40}$$

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