Chapter 2

Fundamentals of Probability

This chapter briefly introduces the fundamentals of probability theory.

2.1 What are probabilities?

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 if the coin is flipped, the probability of it 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 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 \( p \) 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 [?](?), a beautiful, short paper.

### 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, if you toss two coins, the sample space is

\[ \Omega = \{hh, ht, th, hh\} \]

where h is Heads and t is Tails. Sample spaces can be finite, countably infinite (e.g., the set of integers), 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.

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

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

A **probability space** \( P \) on \( \Omega \) is a function from events in \( \Omega \) to real numbers such that the following three properties hold:

1. \( P(\Omega) = 1 \).
2. \( P(E) \geq 0 \) for all \( E \subset \Omega \).
3. If \( E_1 \) and \( E_2 \) are disjoint, then \( P(E_1 \cup E_2) = P(E_1) + P(E_2) \).
2.4 Conditional Probability 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 an 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” (?, ?, ?). Suppose that among transitive sentences in a corpus of historical English, the frequency distribution of object position and pronominality is as follows:

<table>
<thead>
<tr>
<th></th>
<th>Pronoun</th>
<th>Not Pronoun</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object Preverbal</td>
<td>0.224</td>
<td>0.655</td>
</tr>
<tr>
<td>Object Postverbal</td>
<td>0.014</td>
<td>0.107</td>
</tr>
</tbody>
</table>

For the moment, we will interpret these frequencies directly as probabilities. (More on this in Chapter 3.) What is the conditional probability of pronominality given that an object is postverbal?

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

$$\frac{0.014}{0.121} = 0.116$$

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” (?, ?, inter alia).
\begin{equation}
\begin{aligned}
P(\text{Pronoun}|\text{Postverbal}) &= \frac{P(\text{Postverbal} \cap \text{Pronoun})}{P(\text{Postverbal})} \\
&= \frac{0.014}{0.014 + 0.107} = 0.116
\end{aligned}
\end{equation}

2.4.1 (Conditional) Independence

Events \( A \) and \( B \) are said to be CONDITIONALLY INDEPENDENT GIVEn \( C \) if

\[ P(A \cap B|C) = P(A|C)P(B|C) \]

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 a state \( C \) of “not knowing anything at all” \( (C = \emptyset) \).

It’s crucial to keep in mind that if \( A \) and \( B \) are conditionally independent given \( C \), that does not guarantee they will be conditionally independent given some other set of knowledge \( C' \).

2.5 Discrete Random Variables and Probability Densities

A discrete random variable \( X \) is literally a function from the sample space \( \Omega \) of a probability space to a finite, or countably infinite, set of real numbers \((\mathbb{R})\).\(^2\) Together with the function \( P \) mapping elements \( \omega \in \Omega \) to probabilities, a random variable determines a probability density \( P(X(\omega)) \), or \( P(X) \) for short, over the real numbers.

The relationship between the sample space \( \Omega \), a probability space \( P \) on \( \Omega \), and a discrete random variable \( X \) on \( \Omega \) can be a bit subtle, so we’ll illustrate it by returning to our example of tossing two coins. Once again, the sample space is \( \Omega = \{hh, ht, th, hh\} \). Consider the function \( X \) that maps every possible outcome of the coin flipping—that is, every point in the sample space—to the total number of heads obtained in that outcome. Suppose further that both coins are fair, so that for each point \( \omega \) in the sample space we have \( P(\omega) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \).

The total number of heads is a random variable \( X \), and we can make a table of the relationship between \( \omega \in \Omega \), \( X(\omega) \), and \( P(X) \):

\[^2\text{A set } S \text{ is COUNTABLY INFINITE if a one-to-one mapping exists between } S \text{ and the integers } 0, 1, 2, \ldots.\]
Notice that the random variable $X$ serves to partition the sample space into equivalence classes: for each possible real number $y$ mapped to by $X$, all elements of $\Omega$ mapped to $y$ are in that equivalence class. By the axioms of probability (Section 2.3), $P(y)$ is simply the sum of the probability of all elements in $y$’s equivalence class.

### 2.6 Continuous random variables and probability densities

Continuous probability densities (CPDs) are those over random variables whose values can fall anywhere in one or more continua on the real number line. For example, the amount of time that an infant has lived before it hears a parasitic gap in its native-language environment would be naturally modeled as a continuous probability distribution.

With discrete probability distributions, the probability density function (PDF, often called the PROBABILITY MASS FUNCTION for discrete random variables) assigned a non-zero probability to points in the sample space. That is, for such a probability space you could “put your finger” on a point in the sample space and there would be non-zero probability that the outcome of the r.v. would actually fall on that point. For cpd’s, this doesn’t make any sense. Instead, the pdf is a true density in this case, in the same way as a point on a physical object in pre-atomic physics doesn’t have any mass, only density—only volumes of space have mass.

As a result, the cumulative distribution function (CDF; $P(X \leq x)$ is of primary interest for cpd’s, and its relationship with the probability density function $p(x)$ is defined through an integral:

$$P(X \leq x) = \int_{-\infty}^{x} p(x)$$

We then become interested in the probability that the outcome of an r.v. will fall into a region $[x, y]$ of the real number line, and this is defined as:

$$P(x \leq X \leq y) = \int_{x}^{y} p(x) = P(X \leq y) - P(X \leq x)$$

Note that the notation $f(x)$ is often used instead of $p(x)$.
2.7 Expected values and variance

We now turn to two fundamental quantities of probability distributions: **expected value** and **variance**.

### 2.7.1 Expected value

The expected value of a random variable $X$, denoted $E(X)$ or $E[X]$, is also known as the **mean**. For a discrete random variable $X$ under probability distribution $P$, it's defined as

$$E(X) = \sum x_i P(x_i)$$

For a continuous random variable $X$ under cpd $p$, it's defined as

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

What is the mean of a binomially-distributed r.v. with parameters $n, p$? What about a uniformly-distributed r.v. on $[a, b]$?

Sometimes the expected value is denoted by the Greek letter $\mu$, “mu”.

#### Linearity of the expectation

Linearity of the expectation 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 $E(Y) = \sum_i E(X_i)$.

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

$$E(Y) = a + \sum_i b_i E(X_i)$$  \hspace{1cm} (2.3)

This is incredibly convenient. For example, it is intuitively obvious that the mean of a binomially distributed r.v. $Y$ with parameters $n, p$ is $pn$. 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** $X_i$. (A Bernoulli random variable is a single coin toss with probability of success $p$.) The mean of each $X_i$ is trivially $p$, so we have:

$$E(Y) = \sum_i^n E(X_i) \sum_i^n p = pn$$
2.7.2 Variance

The variance is a measure of how broadly distributed the r.v. tends to be. It’s defined in terms of the expected value:

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

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

2.8 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

\footnote{Technically, the definition of the multivariate cpd is then}

$$F(x_1, \cdots, x_n) \overset{\text{def}}{=} P(X_1 \leq x_1, \cdots, X_N \leq x_n) = \sum_{\bar{x}\leq (x_1,\cdots,x_N)} p(\bar{x}) \quad \text{[Discrete]} \quad (2.4)$$

$$F(x_1, \cdots, x_n) \overset{\text{def}}{=} P(X_1 \leq x_1, \cdots, X_N \leq x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_N} p(\bar{x})d\bar{x} \quad \text{[Continuous]} \quad (2.5)$$
Figure 2.2: The probability of the formants of a vowel landing in the grey rectangle can be calculated using the joint cumulative distribution function.

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

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:

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

and visualized in Figure 2.2 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)

### 2.8.1 Multinomial distributions as jpd’s

We touched on the multinomial distribution very briefly in a previous lecture. To refresh your memory, a multinomial can be thought of as a random set of \( n \) outcomes into \( r \) distinct classes—that is, as \( n \) rolls of an \( r \)-sided die where the tabulated outcome is the number of rolls that came up in each of the \( r \) classes. This outcome can be represented as a set of random variables \( X_1, \cdots, X_r \), or equivalently an \( r \)-dimensional real-valued vector. The \( r \)-class multinomial distribution is characterized by \( r-1 \) parameters, \( p_1, p_2, \cdots, p_{r-1} \), which are the probabilities of each die roll coming out as each class. The probability of the die roll coming out in the \( r \)th class is \( 1 - \sum_{i=1}^{r-1} p_i \), which is sometimes called \( p_r \) but is not a true parameter of the model. The probability mass function looks like this:
\[ p(n_1, \ldots, n_r) = \binom{n}{n_1 \ldots n_r} \prod_{i=1}^{r} p_i \]

### 2.9 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 \(X, Y\):

\[
P(X = x) = \sum_y P(x, y) = \sum_y P(X = x | Y = y) P(y)
\]

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**.

### 2.10 Covariance

The **covariance** between two random variables \(X\) and \(Y\) is defined as follows:

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

Simple example:

<table>
<thead>
<tr>
<th>Coding for (X)</th>
<th>Coding for (Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Object</strong></td>
<td><strong>Pronoun</strong></td>
</tr>
<tr>
<td>0</td>
<td>0.224</td>
</tr>
<tr>
<td>1</td>
<td>0.014</td>
</tr>
</tbody>
</table>

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 result, we have \(\mu_X = 0.121\), \(\mu_Y = 0.762\). The covariance between the two is:

\(^4\)or integral in the continuous case
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:

```r
> 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:

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

The covariance is essentially zero now.

An important aside: the variance of a random variable $X$ is just its covariance with itself:

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

### 2.10.1 Covariance and scaling random variables

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

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

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

---

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

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

$$= E[(bX - bE(X))(Y - E(Y))]$$

$$= E[b(X - E(X))(Y - E(Y))]$$

$$= bE[(X - E(X))(Y - E(Y))]$$

$$= b \text{Cov}(X, Y)$$
2.10.2 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 common to use the **correlation** \( \rho \) as a standardized form of covariance:

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

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