PROBABILITIES

We review the three useful properties of conditional probability that were stated in Lecture 15: the Chain Rule, Bayes’ Rule, and the Rule of Total Probability.

**Proposition 16.1 (The Chain Rule)** *For any events* $A_1, \ldots, A_n$ *we have*

$$\Pr[A_1 \cap A_2 \cap A_3 \cdots \cap A_n] = \Pr[A_1] \cdot \Pr[A_2 \mid A_1] \cdot \Pr[A_3 \mid A_1 \cap A_2] \cdots \cdots \Pr[A_n \mid A_1 \cap \cdots \cap A_{n-1}]$$

**Proof:** By induction on $n$. We omit the details but the key observation in the induction step is that

$$\Pr[A_1 \cap \cdots \cap A_k \cap A_{k+1}] = \Pr[A_1 \cap \cdots \cap A_k] \cdot \Pr[A_{k+1} \mid A_1 \cap \cdots \cap A_k]$$

Alternatively, we can show that the RHS is telescoping product. ■

**Proposition 16.2 (Bayes’ Rule)**

$$\Pr[A \mid B] = \frac{\Pr[A] \Pr[B \mid A]}{\Pr[B]}$$

The proof is immediate from the definition of conditional probability.

**Proposition 16.3 (The Rule of Total Probability)** *Let* $A_1, \ldots, A_n$ *be a set of events of non-zero probability in* $(\Omega, \Pr)$ *such that:

- $A_1, \ldots, A_n$ *are pairwise disjoint, and
- $A_1 \cup \cdots \cup A_n = \Omega$

*(we say that* $A_1, \ldots, A_n$ *form a partition of* $\Omega$). *Then, for any event* $E$

$$\Pr[E] = \sum_{i=1}^{n} \Pr[E \mid A_i] \Pr[A_i]$$

*In particular, if* $\Pr[A] \neq 0$ *then* $\Pr[E] = \Pr[E \mid A] \Pr[A] + \Pr[E \mid \overline{A}] \Pr[\overline{A}]$. 


Proof: We only show the case for $A$, $\overline{A}$. (The general case is proved by induction.) Note that $E \cap A$ and $E \cap \overline{A}$ are disjoint and verify that $(E \cap A) \cup (E \cap \overline{A}) = E$. Then the equality follows by the definition of conditional probability.

Random variables

Definition 16.4 Let $(\Omega, \Pr)$ be a (finite) probability space. A random variable on this space is a function $X : \Omega \to \mathbb{R}$. It is customary to denote with $x$ (and such) the real values that $X$ returns. We denote $\text{Val}(X) = \{x \in \mathbb{R} | \exists w \in \Omega \ X(w) = x\}$. Note that $\text{Val}(X)$ is also a finite set.

We denote with $X = x$ the event \{\(w \in \Omega \mid X(w) = x\)\}. The probabilities of the events of the form $X = x$ are what we will call the distribution of the random variable $X$.

Random variables play an essential, in fact, indispensable, role in statistics and machine learning. According to their definition, they are introduced to model the measuring of various quantities that depend on random phenomena/experiments/processes.

Example 16.5 We roll a fair die. What is the distribution of the random variable $D$ that returns the number shown by the die? We roll two fair dice. What is the distribution of the random variable $S$ which returns the sum of the numbers the two dice show?

Solution: $\text{Val}(D) = [1..6]$ and $\text{Val}(S) = [2..12]$. The graphs of the distributions of $D$ and $S$ are in Figure 1.

Observe that for the distribution of $S$ we have
\[
\frac{1}{36} + \frac{1}{18} + \frac{1}{12} + \frac{1}{9} + \frac{5}{36} + \frac{1}{6} + \frac{5}{36} + \frac{1}{9} + \frac{1}{12} + \frac{1}{18} + \frac{1}{36} = 1
\]

This is true for any random variable:

Proposition 16.6
\[
\sum_{x \in \text{Val}(X)} \Pr[X = x] = 1
\]

1Therefore, a random variable is neither random nor a variable! But that’s what everybody calls these mathematical objects.

2Strictly speaking, the (cumulative) distribution function of $X$ is $F : \mathbb{R} \to [0,1]$ where $F(x) = \Pr[X \leq x]$. Here we denote with $X \leq x$ the event \{\(w \in \Omega \mid X(w) \leq x\)\}. Meanwhile the function $f : \mathbb{R} \to [0,1]$ where $f(x) = \Pr[X = x]$ is called the probability mass function of $X$ (both functions are defined to be 0 on values outside of $\text{Val}(X)$). The “(cumulative) distribution” terminology originated with infinite probability spaces where instead of the probability mass a function called “probability density” is preferred. For finite probability spaces we can easily go back and forth between mass and cumulative distribution and we will just call “distribution” the former.
Problem 16.7 \textit{Probabilities add up to 1 for both the probability distribution that is part of a probability space and for the probability distribution associated with a random variable. This is not accidental. For any random variable $X$ the values that $X$ returns form a probability space (describe it!). Further, a random variable is a mapping between two probability spaces. More generally, let $(\Omega_1, \Pr_1)$ and $(\Omega_2, \Pr_2)$ be two probability spaces and let $f : \Omega_1 \rightarrow \Omega_2$ be a function. State a property that captures the intuitive idea “the two spaces are compatible via $f$”. Explain how you can use this idea to “clean up” probability spaces by eliminating outcomes of probability 0 and outcomes of probability 1. Also explain how this idea helps one understand the relationship between the two alternative formalizations of “marbles from urns”.

In two previous occasions we have discussed what were, essentially, random variables, without naming them as such. We can now add a bit more to our terminology.

Definition 16.8 \textit{Recall that a Bernoulli trial is an experiment with two random outcomes, success (with probability $p$) and failure (with probability $1 - p$). We often consider the random variable $X$ on this space with $X(\text{success}) = 1$ and $X(\text{failure}) = 0$. This random variable is called a Bernoulli random variable and its distribution, $\Pr[X = 1] = p, \Pr[X = 0] = 1 - p$, is called a Bernoulli distribution.}

Consider also $n$ iid Bernoulli trials each with probability of success $p$, and the random variable $B$ that returns the number of successes $\text{Val}(B) = [0..n]$ whose distribution is given by

$$\Pr[B = k] = \binom{n}{k} p^k (1 - p)^{n-k}$$
For obvious reasons, $B$ is called a **binomial** random variable and its distribution is called the **binomial distribution**. An important special case is when $p = 1/2$ and thus $\Pr[B = k] = (1/2^n)(n\choose k)$.

**Expectation**

We are often interested in the average, or mean, value returned by a random variable. However, such an average should be **weighted** by the probability distribution (recall that the probabilities sum up to 1). It turns out that there are two candidates for such a weighted average. Happily, the next proposition shows that they give the same answer!

**Proposition 16.9** For a random variable $X$ defined on $(\Omega, \Pr)$ we have

$$
\sum_{x \in \text{Val}(X)} x \cdot \Pr[X = x] = \sum_{w \in \Omega} X(w) \cdot \Pr[w]
$$

**Proof:**

$$
\sum_{x \in \text{Val}(X)} x \cdot \Pr[X = x] = \sum_{x \in \text{Val}(X)} x \cdot \left[ \sum_{w \in [X=x]} \Pr[w] \right]
$$

$$
= \sum_{x \in \text{Val}(X)} \sum_{w \in [X=x]} x \cdot \Pr[w]
$$

$$
= \sum_{x \in \text{Val}(X)} \sum_{w \in [X=x]} X(w) \cdot \Pr[w]
$$

$$
= \sum_{w \in \Omega} X(w) \cdot \Pr[w]
$$

where we have used two facts: (1) $w \in [X = x]$ iff $X(w) = x$, and (2) the events $[X = x]$, $x \in \text{Val}(X)$ are pairwise disjoint and their union equals $\Omega$. ■

**Definition 16.10** The **expectation** (mean) of an random variable, notation $\mathbb{E}[X]$, is defined by one of the two sums shown equal in the previous proposition.

Remarkably, the first sum in the proposition can be computed without knowledge of the probability space on which $X$ is defined: we just use the values that $X$ returns and the distribution of $X$. Consequently, when we talk about the “mean or expectation of a distribution”, this is understood to be the expectation of an random variable with that given distribution.

On uniform probability spaces the expectation is actually the average of the values returned by an random variable as illustrated in the following example.
Example 16.11 Compute $E[D]$ where $D$ is the number shown by a fair die (the random variable from Example 16.5).

Solution: We use the first sum in Proposition 16.9.

$$1 \cdot (1/6) + 2 \cdot (1/6) + 3 \cdot (1/6) + 4 \cdot (1/6) + 5 \cdot (1/6) + 6 \cdot (1/6) = (1 + 2 + 3 + 4 + 5 + 6)/6 = 3.5$$

Example 16.12 We roll two fair dice. What is the expectation of the random variable $S$ which returns the sum of the numbers the two dice show? (See Example 16.5.)

Solution: Refer to the graph of the distribution of $S$ in Figure 1.

$$E[S] = 2 \cdot (1/36) + 3 \cdot (2/36) + 4 \cdot (3/36) + 5 \cdot (4/36) + 6 \cdot (5/36) + 7 \cdot (6/36) + 8 \cdot (5/36) + 9 \cdot (4/36) + 10 \cdot (3/36) + 11 \cdot (2/36) + 12 \cdot (1/36)$$

$$= 252/36 = 7$$

Linearity of expectation (see below) will give us an easier calculation.

Example 16.13 Let $X$ be a Bernoulli random variable with probability of success $p$. Compute its expectation.

Solution: $E[X] = 1 \cdot p + 0 \cdot (1 - p) = p$.

It would be now the turn of expectation for binomial random variables. However, linearity of expectation, shown next, will give us a much easier calculation than the direct evaluation of the sum given by the definition, see Example 16.17 below.

Operations on random variables: sum, scalar multiplication, product

Let $(\Omega, \Pr)$ be a probability space and $X, Y$ two random variables on this space. Then $X + Y$ is the random variable on the same space defined by $[X + Y](w) = X(w) + Y(w)$ for all $w \in \Omega$. Let $c \in \mathbb{R}$. Then $cX$ is the random variable on the same space defined by $(cX)(w) = c \cdot X(w)$. Similarly we can define $XY$, and, of course, we can define the sum, or product, of three or more random variables.

The following is perhaps the property of probability that is most often used in reasoning about randomness in algorithms, as it can greatly simplify reasoning and calculations.
Proposition 16.14 (Linearity of expectation) Let $X_1, \ldots, X_n$ be random variables on the same probability space $(\Omega, \Pr)$ and let $c_1, \ldots, c_n \in \mathbb{R}$. (No other assumptions needed!) Then:

$$E[c_1 X_1 + \cdots + c_n X_n] = c_1 E[X_1] + \cdots + c_n E[X_n]$$

Proof: Interestingly, this proof is easy if we use the second sum from Proposition 16.9 (and quite messy if we use the first one!).

$$E[\sum_{i=1}^{n} c_i X_i] = \sum_{w \in \Omega} \left( \sum_{i=1}^{n} c_i X_i(w) \right) \cdot \Pr[w]$$

$$= \sum_{w \in \Omega} \left( \sum_{i=1}^{n} c_i X_i(w) \right) \cdot \Pr[w]$$

$$= \sum_{i=1}^{n} c_i \left( \sum_{w \in \Omega} X_i(w) \cdot \Pr[w] \right)$$

$$= \sum_{i=1}^{n} c_i E[X_i]$$

Using this, the solution to Example 16.12 follows immediately since we already know from Example 16.11 the expectation of the value shown by each die: $E[S] = E[D_1 + D_2] = E[D_1] + E[D_2] = 3.5 + 3.5 = 7$.

Indicator random variables

Linearity of expectation is almost always used for a sum of indicator random variables.

Definition 16.15 Let $A$ be an event in a probability space $(\Omega, \Pr)$. The **indicator** random variable of the event $A$ is defined by

$$I_A(w) = \begin{cases} 1 & \text{if } w \in A \\ 0 & \text{if } w \notin A \end{cases}$$

Note that $I_A$ is a Bernoulli random variable with success probability $\Pr[A]$. As we have shown before, its expectation is $E[I_A] = \Pr[A]$.

Example 16.16 Recall the balls-into-bins model from Lecture 14. We throw $k$ balls into $n$ bins. What is the number of balls that end up in Bin 1, “on average”?

Solution: Let $X$ be the random variable that returns the number of balls that end up in Bin 1. We can express $X$ as a sum of indicator variables:

$$X = I_{L_1} + \cdots + I_{L_k}$$
where \( L_i \) is the event “ball \( i \) ends up in Bin 1” for \( i = 1, \ldots, k \).

Recall from the discussion that introduced the model that \( \Pr[L_i] = 1/n \). By Example 16.13 we have \( \E[I_{L_i}] = 1/n \).

Using linearity of expectation we obtain \( \E[X] = (1/n) + \ldots + (1/n) = k/n \).

Note that \( X \) is a binomial random variable with parameters \( k \) and \( 1/n \). In the next example we consider binomial random variables in general.

\[ \text{Example 16.17} \quad \text{Let } B \text{ be a binomial random variable with parameters } n \text{ and } p. \text{ Compute its expectation.} \]

\[ \text{Solution: The formula yields:} \]

\[ \sum_{k=0}^{n} k \cdot \binom{n}{k} p^{k} (1-p)^{n-k} \]

The challenge is to find a simple form for this sum. This is doable, there is even a trick that takes derivatives on both sides of the binomial theorem, but the use of indicator random variables and of linearity of expectation gives us a much simpler solution. In fact, we have already solved a particular case in Example 16.16.

In the probability space corresponding to \( n \) iid Bernoulli trials (Example 14.14 in Lecture 14) let \( S_k \) be the event “the \( k \)'th trial is a success” for \( k = 1, \ldots, n \) and let \( I_{S_k} \) be its indicator random variable. From Example 16.13 we have \( \E[I_{S_k}] = \Pr[S_k] \) and we compute \( \Pr[S_k] \) as follows.

Each outcome in \( S_k \) has success for trial \( k \) and \( i \) other successes where \( i \) ranges from 0 to \( n-1 \). Therefore it has probability \( p \cdot p^i \cdot (1-p)^{n-1-i} \). There are \( \binom{n-1}{i} \) outcomes with success in trial \( k \) and \( i \) other successes. Therefore, using the binomial theorem,

\[ \Pr[S_k] = \sum_{i=0}^{n-1} \binom{n-1}{i} p \cdot p^i \cdot (1-p)^{n-1-i} = p \sum_{i=0}^{n-1} \binom{n-1}{i} p^i \cdot (1-p)^{n-1-i} = p(p + 1 - p)^{n-1} = p \]

which corresponds to the intuition about \( S_k \).

Now \( B = I_{S_1} + \cdots + I_{S_n} \) (why?), therefore \( \E[B] \) can be computed using linearity of expectation:

\[ \E[B] = \sum_{k=1}^{n} \E[X_k] = \sum_{k=1}^{n} p = np \]

\[ \text{Example 16.18} \quad \text{Consider again } n \text{ iid Bernoulli trials, each with probability of success } p. \text{ Let } C \text{ be the random variable that returns the number of successes which are followed immediately (in the next trial) by failure. Compute the expectation of } C. \]

7
Solution: Trying to compute this from one of the definitions of expectation is very complicated. However, notice that $C = I_{A_1} + \cdots + I_{A_{n-1}}$ where $A_i$ is the indicator variable of the event “success in trial $i$ and failure in trial $i+1$” for $i = 1, \ldots, n-1$.

As always with applying linearity of expectation for a sum of indicator variables we need $\Pr[A_i]$ for $i = 1, \ldots, n-1$. Note that $A_i = S_i \cap F_{i+1}$ where $S_i$ is the event “success in trial $i$” and $F_{i+1}$ is the event “failure in trial $i+1$”. Since the trials are independent, these two events are independent (recall again Example 14.14 in Lecture 14) and since $\Pr[S_i] = p$ and $\Pr[F_{i+1}] = 1 - p$ (see the solution to Example 16.17) we have $\Pr[A_i] = p(1 - p)$ and therefore $E[I_{A_i}] = p(1 - p)$.

By linearity of expectation, it follows that $E[C] = (n - 1)p(1 - p)$.

There is another interesting angle to this solution. The events $A_1, \ldots, A_{n-1}$ are, in general, not mutually independent. In fact, they are not even pairwise independent because pairs of consecutive events are disjoint: every outcome in $A_1$ has failure in the second trial and thus cannot be in $A_2$! After we give an appropriate definition (in Lecture 17) we shall see that their indicator random variables $I_{A_1}, \ldots, I_{A_{n-1}}$ are not independent either. Lack of independence can significantly complicate probability calculations. But this does not matter here because linearity of expectation (Proposition 16.14) does not require independence! ■

Next is another example in which linearity of expectation conquers events that are not independent in a complicated probability space.

Example 16.19 (The Hat Check Problem)³ A hapless cloakroom attendant randomly returns their checked hats to $n$ associates of Al Capone. How many of the hats end up in the hands of their owner, “on average”?

Equivalently, let $S$ be a set of $n$ distinct real numbers. I try to sort them by producing a random permutation of the elements of $S$. My brilliant intuition tells me that this cannot be too bad, can it? I should get maybe half of the numbers in their place? Am I right?

Solution: Recall Example 13.10 in Lecture 13. We have a uniform probability space whose outcomes are all permutations of $S$ so each permutation has probability $1/n!$.

We are interested in the expectation of the random variable $X$ which returns the number of elements of $S$ that end up in their properly sorted place in a random permutation. To be more precise, we can assume, w.l.o.g., that the elements of $S$ are $a_1 < a_2 < \cdots < a_n$. Then, for any random permutation $w$

$$X(w) = |\{k \in [1..n] \mid a_k \text{ is in position } k \text{ in } w\}|$$

and we wish to compute $E[X]$. Doing this directly from one of the definitions of expectation is complicated, but linearity of expectation gives us, again, an easy way to do it.

³Some probabilists call this the “Matching Problem” but the term “matching” is heavily used elsewhere in computer science.
Let $E_k$ be the event “$a_k$ occurs in position $k$ (in a random permutation)”. In Example 13.10 (Lecture 13) we have calculated $\Pr[E_k] = 1/n$. Note that as we saw in Example 14.13 (Lecture 14) these events are not even pairwise independent, in general.

We can express $X = X_1 + \cdots + X_n$ where $X_k = I_{E_k}$, the indicator random variable of the event $E_k$. Each of $X_1, \ldots, X_n$ is a Bernoulli random variable with parameter $1/n$ because we computed $\Pr[E_k] = 1/n$ in Example 13.10 (Lecture 13).

Hence we have, by linearity of expectation, $E[X] = n \cdot (1/n) = 1$.

On average, only one element ends up in the right place. Did you guess that it would be that bad? \footnote{\text{By the way, poor cloakroom attendant!}}