1 Linearity of expectation

We will prove the linearity of expectation now:

**Theorem 1** (Linearity of expectation). If $X_1, \ldots, X_n$ are random variables defined on a probability space $(\Omega, P)$, $a_1, \ldots, a_n$ and $b$ are real numbers, and $X = (\sum_{i=1}^n a_i X_i) + b$, we have that

$$E[X] = E[(\sum_{i=1}^n a_i X_i) + b] = (\sum_{i=1}^n a_i E[X_i]) + b.$$ 

We will prove this in two parts. We will first show the additivity of expectation for two variables (it’s easy to make the proof work for more than two variables), and then we will look at how to deal with expectations of random variables with multiplicative and additive constants.

**Lemma 2.** Let $X$ and $Y$ be random variables defined on $(\Omega, P)$. Then

$$E[X + Y] = E[X] + E[Y].$$

**Proof.** We will use the first definition of expectation. Let $Z = X + Y$. Recall that $X, Y, Z$ are all just functions of the from $\Omega \to \mathbb{R}$, and so we have that for every $\omega \in \Omega$, $Z(\omega) = X(\omega) + Y(\omega)$.

Using the definition of $E[Z]$,

$$E[Z] = \sum_{\omega \in \Omega} (X(\omega) + Y(\omega))P(\{\omega\}).$$

We now expand the product to get

$$E[Z] = \sum_{\omega \in \Omega} X(\omega)P(\{\omega\}) + \sum_{\omega \in \Omega} Y(\omega)P(\{\omega\}) = E[X] + E[Y],$$

where there last equation follows from the definition of $E[X]$ and $E[Y]$.

**Lemma 3.** Let $X$ be a random variable defined on a probability space $(\Omega, P)$. Let $c, d \in \mathbb{R}$ be constants. Define the random variable $Y = cX + d$. Then

Proof. We will use the first definition of expectation:

$$\mathbb{E}[Y] = \sum_{\omega \in \Omega} Y(\omega)P(\{\omega\}).$$

Recall that $Y$ and $X$ are basically functions from $\Omega$ to $\mathbb{R}$, and since $Y = cX + d$, for every $\omega \in \Omega$, $Y(\omega) = cX(\omega) + d$. This means that

$$\mathbb{E}[Y] = \sum_{\omega \in \Omega} (cX(\omega) + d)P(\{\omega\}).$$

We can use distributivity to rewrite this as:

$$\mathbb{E}[Y] = \sum_{\omega \in \Omega} (cX(\omega)P(\{\omega\}) + dP(\{\omega\})) = \sum_{\omega \in \Omega} cX(\omega)P(\{\omega\}) + \sum_{\omega \in \Omega} dP(\{\omega\}).$$

Since $c$ and $d$ are constants we can pull them out of the sums:

$$c \left( \sum_{\omega \in \Omega} X(\omega)P(\{\omega\}) \right) + d \left( \sum_{\omega \in \Omega} P(\{\omega\}) \right) = c\mathbb{E}[X] + d.$$

Here we used the fact that $\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega)P(\{\omega\})$ and $\sum_{\omega \in \Omega} P(\{\omega\}) = 1$. □

It’s not hard to combine the two lemmas to prove the theorem of linearity of expectation stated above.

**Question.** Suppose I roll a fair dice 10 times. After each throw, I observe the number rolled, multiply it by 5, and note it down. At the end of all the throws, I add up the numbers I noted down and add 6 to the whole sum. I call this sum $X$. What’s $\mathbb{E}[X]$?

**Proof.** Let $X_1, \ldots, X_{10}$ denote the random variables that are equal to the numbers rolled in the various throws. Then, we have that

$$X = 5X_1 + 5X_2 + \ldots + 5X_{10} + 6 = 5(\sum_{i=1}^{10} X_i) + 6.$$ 

Thus, using linearity of expectation we have

$$\mathbb{E}[X] = \mathbb{E}[5(\sum_{i=1}^{10} X_i) + 6] = \mathbb{E}[5(\sum_{i=1}^{10} X_i)] + 6 = 5(\sum_{i=1}^{10} \mathbb{E}[X_i]) + 6.$$ 

We know that $\mathbb{E}[X_i] = 3.5$ for every $i$, and so $\mathbb{E}[X] = 181$. □

### 1.1 Infinite sample spaces

In many cases, the sample space $\Omega$ can be countably infinite\footnote{It can be uncountably infinite but that is beyond the scope of this course.} and/or the random variables can have a countably infinite range. While we wouldn’t be studying this case in its full generality in
this course, we will see certain instances of experiments which warrant the use of infinite sample spaces or random variables. The examples we will see are all “nice” cases and so whatever we have learnt about probability and random variables for the finite case will apply to them without any change. The only difference is that sometimes when computing the expectation of such countably infinite random variables, we will have to use infinite series to solve them.

Let’s see an example:

**Question**. Suppose we keep tossing a fair coin till we see heads. Let $X$ be the total number of tosses till we see a heads. What is $E[X]$?

**Proof.** Clearly, $\Omega$ in this case is countably infinite: the outcomes are $H$, $TH$, $TTH$, ......., and we have that $P(H) = 1/2$, $P(TH) = 1/4$, $P(TTH) = 1/8$, .......

Let $X_1$ be the random variable that’s 1 if the first flip is a tails and it’s 0 otherwise. Let $X_2$ be the random variable which is 1 if the first and second flips results in tails, and 0 otherwise. Similarly, we can define $X_i$ to be 1 if the first $i$ coin tosses are tails and 0 otherwise. We will define $X_i$ for every natural number $i$ and so we have a countably infinite number of variables. Then, if $X$ is the number of times we flip before we see a heads:

$$X = 1 + X_1 + X_2 + \ldots + (\text{to } \infty).$$

You can convince yourself that the above equation is indeed true: just try to analyze the different cases; the case when you see a heads in the first toss, the case when you see a heads in the second toss, and so on. Then, using linearity of expectation (Let’s assume it works for infinitely many random variables without getting into the details):

$$E[X] = 1 + \sum_{i=1}^{\infty} E[X_i].$$

What is $E[X_i]$? Since $X_i$ is an indicator random variable (for what event? why?), we know that

$$E[X_i] = P(X_i = 1) = P(\text{the first } i \text{ coin tosses result in tails}) = \left( \frac{1}{2} \right)^i.$$

Here the last equation follows from the fact that all the coin tosses are independent. So we have

$$E[X] = 1 + \sum_{i=1}^{\infty} \frac{1}{2^i} = \sum_{i=0}^{\infty} \frac{1}{2^i}.$$

Recall from calculus that for $0 < a < 1$:

$$\sum_{i=0}^{\infty} a^i = \frac{1}{1 - a}.$$

In our case $a = 1/2$, and so we get

$$E[X] = \frac{1}{1 - 1/2} = 2.$$
2 Conditional expectation

Let $X$ be a random variable defined on a probability space $(\Omega, P)$, and let $A$ be an event of the probability space. Then, we can define the expectation of $X$ given that the event $A$ has happened. A natural way to do this is to look at the expression for expectation of $X$ and replace probability with the conditional probability. In particular, recall the second definition of expectation:

$$E[X] = \sum_{a \in \text{Range}(X)} aP(X = a).$$

To define the conditional expectation of $X$, i.e. the expectation conditioned on the event $A$, we simply replace $P(X = a)$ by $P(X = a | A)$, to get

$$E[X | A] = \sum_{a \in \text{Range}(X)} aP(X = a | A).$$

**Question.** Suppose we keep throwing a dice till we see a 6. Let $Y$ be the total number of throws needed, and let $X$ be the total number of ones observed. Compute $E[X | Y = 10]$.

**Proof.** Since we have to find the expectation of $X$ given that $Y = 10$, we can only focus on the first 10 throws. In fact, since we know that the 10th dice must be a 6, we can just focus on the first 9 throws. Let $X_1, \ldots, X_9$ be random variables such that $X_i$ is the indicator random variable for the event that the $i$th throw results in a 1. Then, we have that

$$E[X | Y = 10] = E[\sum_{i=1}^9 X_i | Y = 10].$$

This is because once we are given the information that $Y = 10$, $X$ can be written as $X_1 + X_2 + \ldots + X_9$ (why?). Also to be noted is the fact that linearity of expectation holds even in the case of conditional expectation (the same proof as before will work. One simply has to replace probability by conditional probability). Thus,

$$E[X | Y = 10] = E[\sum_{i=1}^9 X_i | Y = 10] = \sum_{i=1}^9 E[X_i | Y = 10].$$

We now need to compute $E[X_i | Y = 10]$. Let’s just use the definition of conditional expectation to do so:

$$E[X_i | Y = 10] = \sum_{a \in \text{Range}(X_i)} aP(X_i = a | Y = 10) = 1 \cdot P(X_i = 1 | Y = 10) + 0 \cdot P(X_i = 0 | Y = 10).$$

$$= P(X_i = 1 | Y = 10) = \frac{P(X_i = 1 \cap Y = 10)}{P(Y = 10)}.$$

\(\square\)
It’s not hard to compute that \( P(Y = 10) = \frac{5}{6} \frac{9}{6} \), and \( P(X_i = 1 \cap Y = 10) = \frac{(5/6)^8}{(5/6)^9} = \frac{1}{5} \).

This means that
\[
\mathbb{E}[X|Y = 10] = \sum_{i=1}^{9} \frac{1}{5} = \frac{9}{5}.
\]

### 3 Law of total expectation

The way we can compute a probability of an event \( B \) by computing the probability of \( B \) happening conditioned on disjoint events \( A_1, \ldots, A_n \) that form a partition of \( \Omega \) i.e. compute \( P(B|A_1), P(B|A_2), \ldots, \) and then combining all these cases using the law of total probability, i.e. \( P(B) = \sum_{i=1}^{n} P(A_i)P(B|A_i) \), we can also do the same for expectation.

**Theorem 4** (Law of total expectation). Consider a probability space \((\Omega, P)\) and let \( A_1, \ldots A_n \) be disjoint nonempty events in it that form a partition of \( \Omega \). Then,
\[
\mathbb{E}[X] = \sum_{i=1}^{n} P(A_i)\mathbb{E}[X|A_i].
\]

**Proof.** The proof follows from the law of total probability and the second definition of expectation:
\[
\mathbb{E}[X] = \sum_{k \in \text{Range}(X)} kP(X = k).
\]

Using law of total probability, we can write
\[
P(X = k) = \sum_{i=1}^{n} P(A_i)P(X = k|A_i).
\]

Substituting this back into the first equation, we get
\[
\mathbb{E}[X] = \sum_{k \in \text{Range}(X)} k \left( \sum_{i=1}^{n} P(A_i)P(X = k|A_i) \right) = \sum_{k \in \text{Range}(X)} \sum_{i=1}^{n} kP(A_i)P(X = k|A_i).
\]

We can now switch the sums to get
\[
\mathbb{E}[X] = \sum_{i=1}^{n} P(A_i) \left( \sum_{k \in \text{Range}(X)} kP(X = k|A_i) \right).
\]

But the expression inside the paranthesis is just \( \mathbb{E}[X|A_i] \), and so we get
\[
\mathbb{E}[X] = \sum_{i=1}^{n} P(A_i)\mathbb{E}[X|A_i].
\]

\qed
**Question.** Suppose we keep tossing a biased coin (the coin lands heads with probability $p$ and tails with probability $1 - p$) till we see a heads. On average, how many times will we have to toss the coin?

**Proof.** Let $X$ be the number of times the coin is tossed till a heads is observed. Let $A$ be the event that we observe a heads in the very first coin toss. Then using the law of total expectation:

$$E[X] = P(A)E[X|A] + P(A^c)E[X|A^c].$$

Here $P(A) = p$, $P(A^c) = 1 - p$, $E[X|A] = 1$ (why?). How about $E[X|A^c]$? Notice that if the first coin toss is not heads, then starting from the second coin toss, we will need the same number of coin tosses on average to see a heads as we would if starting from the first coin toss (Think about this - it’s important!). Basically, if our first toss results in tails, it’s almost like forgetting that this happened and starting all over again, and so

$$E[X|A^c] = 1 + E[X].$$

The one is added to the RHS to account for the first coin toss that resulted in tails. Plugging in these values in the expression for $E[X]$, we get

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

Let $z = E[X]$. We now have a linear equation in the variable $z$, and we can solve for it:

$$z = p + (1 - p)(1 + z) \implies z = p + 1 - p + z(1 - p)$$

$$\implies pz = 1 \implies z = \frac{1}{p}$$

Thus, $E[X] = \frac{1}{p}$. 

\[\square\]

### 4 Probability mass function

Before we begin let’s set some things straight. Up till now, when we have been talking about probability in the context of probability spaces $(\Omega, P)$, we have been calling them “probability distributions”. While it’s okay use to this term (and some people do), (I think) the more widely used term is “probability measure”. I did not want to introduce this in the beginning to avoid confusion/intimidation.

Formally, with an experiment we associate a sample space $\Omega$ and then we define a probability measure $P$ on $\Omega$ (what we have been calling a probability distribution so far). A probability measure is basically a function $P : 2^\Omega \to \mathbb{R}$, a function that assigns a “probability” with every event.

Having set the record straight, we can now define what we mean by a probability mass function.

**Definition 5** (Probability mass function of a random variable). Let $X$ be a random variable defined on a probability space $(\Omega, P)$ (remember $P$ is called a probability measure on $\Omega$)). The probability mass function $f_X : \text{Range}(X) \to \mathbb{R}$ is a function that assigns to every value in the range of $X$ the probability of $X$ being equal to that value, i.e. for every $a \in \text{Range}(X)$,

$$f_X(a) = P(X = a).$$
So basically, \( f_X(a) \) is just a fancy way of saying \( P(X = a) \), for all we care\(^2\).

In some sense, the probability mass function is like providing a table of probabilities for all the values in the range of \( X \). For example, consider the experiment of tossing a coin three times, and let \( X \) be the number of heads observed. Then, \( \text{Range}(X) = \{0, 1, 2, 3\} \), and we have

\[
\begin{align*}
 f_X(0) &= P(X = 0) = \frac{1}{8} \\
 f_X(1) &= P(X = 1) = \frac{3}{8} \\
 f_X(2) &= P(X = 2) = \frac{3}{8} \\
 f_X(3) &= P(X = 3) = \frac{1}{8}
\end{align*}
\]

To see interesting properties of the probability mass function, you basically have to just look for interesting properties of \( P(X = a) \). We already did this lecture 11.

## 5 Probability distributions

A random variable \( X \) defined on some probability space \((\Omega, P)\) along with its probability mass function \( f_X \) is called a *probability distribution* (Remember, we decided to call \( P \) a probability measure from now on). There are some common probability distributions that arise quite often when modelling experiments and solving problems, and so it’s good to be aware of them.

### 5.1 Bernoulli distribution

A Bernoulli distribution with parameter \( p \) is a random variable \( X \) with range \( \{0, 1\} \) defined on a probability space \((\Omega, P)\), along with its probability mass function \( f_X \) defined as

\[
\begin{align*}
 f_X(1) &= P(X = 1) = p \\
 f_X(0) &= P(X = 0) = 1 - p.
\end{align*}
\]

We typically call \( X \) a *Bernoulli random variable*, or say that \( X \) is distributed according to the Bernoulli distribution. Common situations where Bernoulli random variables/distributions arise are coin tosses and indicator random variables.

The expected value of a Bernoulli random variable is

\[
\mathbb{E}[X] = 1 \cdot P(X = 1) + 0 \cdot P(X = 0) = P(X = 1) = f_X(1) = p.
\]

\(^2\)In the case of continuous random functions, the counterpart of the probability mass function stops being just a fancy alternative to \( P(X = a) \), as you will see if you take a course on probability at some point.
5.2 Geometric distribution

A geometric distribution with parameter $p$ is a random variable $X$ with range $\{1, 2, 3, 4, \ldots \}$ defined on an infinite probability space $(\Omega, P)$, along with its probability mass function $f_X$ defined as

$$f_X(k) = P(X = k) = (1 - p)^{k-1}p$$

for $k \in \text{Range}(X) = \{1, 2, 3, 4, \ldots \}$. A good way to think of geometric distributions is the experiment of tossing a biased coin (which has probability of heads being equal to $p$, and tails $1 - p$) till we observe heads, and thinking of $X$ as the number of tosses. We saw in section 3 (see the example there), that if $X$ is a geometric random variable with parameter $p$, then $E[X] = \frac{1}{p}$.

5.3 Binomial distribution

A binomial distribution with parameters $n$ and $p$ is a random variable $X$ with range $\{0, 1, \ldots, n\}$ defined on a probability space $(\Omega, P)$, along with its probability mass function $f_X$ defined as follows:

$$f_X(k) = P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

for $k \in \text{Range}(X) = \{0, 1, \ldots, n\}$. A good way to think of the binomial distribution is to think of tossing a biased coin (which turns up heads with probability $p$ and tails with $1 - p$) $n$ times and thinking of $X$ as the number of heads observed. Suppose we define indicator (and so Bernoulli) random variables $X_1, \ldots, X_n$ such that $X_i$ is 1 if and only if the $i$th coin toss results in heads, then clearly

$$X = X_1 + \ldots + X_n.$$

Thus, a binomial random variable can be thought of as the sum of $n$ independent Bernoulli random variables, each having parameter $p$. This also means that

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

using linearity of expectation, and since each $E[X_i] = p$ (since they are all Bernoulli), we get that for a binomial distribution the expected value is $np$, i.e. $E[X] = np$.

To see why $P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$, let’s see what the sample space and probability measure are in this case. Let’s represent the outcomes using binary strings of length $n$, where a 1 in the $i$th position indicates that the $i$th coin toss results in heads, a 0 indicates that the $i$th coin toss results in tails. Thus, $\Omega$ is the set of all binary strings of length $n$.

How is the probability measure $P$ on $\Omega$ defined? For an outcome (binary string) $\omega \in \Omega$, if $\omega$ has $k$ ones and $n - k$ zeros, since the probability of heads is $p$ and of tails is $1 - p$, and all the tosses are independent, we have that

$$P(\{\omega\}) = p^k (1 - p)^{n-k}.$$ 

(why is the above true? Try to use the chain rule for independent events if you can’t see why!) The event $[X = k]$, i.e. seeing exactly $k$ heads, consists of all binary strings (outcomes) with $k$ ones and $n - k$ zeros. Since there are $\binom{n}{k}$ such outcomes, and they are all disjoint (outcomes are basically
atomic events, i.e. events that are singleton sets, and so are always disjoint from each other!) and each outcome has probability $p^k(1 - p)^{n-k}$, using the sum rule, we can observe that

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}.$$