7 Random Variables

In this section, we introduce random variables and their expected values. An important property is the linearity of expectations, which offers a shortcut to many results.

**Expectation.** A random variable is a real-value function on the sample space, $X : \Omega \to \mathbb{R}$. An example is the total number of dots at rolling two dice, or the number of heads in a sequence of ten coin flips. The function that assigns to each $x_i \in \mathbb{R}$ the probability that $X = x_i$ is the distribution function of $X$, denoted as $f : \mathbb{R} \to [0, 1]$; see Figure 8. More formally, $f(x_i) = P(A)$, where $A = X^{-1}(x_i)$. The expected value of the random variable is 

$$E(X) = \sum_i x_i P(X = x_i).$$

![Figure 8: The distribution function of a random variable is constructed by mapping a real number, $x_i$, to the probability of the event that the random variable takes on the value $x_i$.](image)

As an example, consider the Bernoulli trial process in which $X$ counts the successes in a sequence of $n$ trials, that is, $P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$. The corresponding distribution function maps $k$ to the probability of having $k$ successes, that is, $f(k) = \binom{n}{k} p^k (1 - p)^{n-k}$. We get the expected number of successes by summing over all $k$.

$$E(X) = \sum_{k=0}^{n} k f(k)$$

$$= \sum_{k=0}^{n} k \binom{n}{k} p^k (1 - p)^{n-k}$$

$$= np \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} (1 - p)^{n-k}$$

$$= np \sum_{k=0}^{n-1} \binom{n-1}{k} p^k (1 - p)^{n-k-1}.$$ 

The sum in the last line is equal to $(p + (1-p))^{n-1} = 1$. Hence, the expected number of successes is $E(X) = np$.

**Linearity of expectation.** Note that the expected value of $X$ can also be obtained by summing over all possible outcomes, that is,

$$E(X) = \sum_{s \in \Omega} X(s)P(s).$$

This leads to an easier way of computing the expected value. To this end, we exploit the following important property of expectations.

**Linearity of Expectation.** Let $X, Y : \Omega \to \mathbb{R}$ be two random variables. Then

(i) $E(X + Y) = E(X) + E(Y)$;
(ii) $E(cX) = cE(X)$, for every real number $c$.

The proof should be obvious. Is it? We use the property to recompute the expected number of successes in a Bernoulli trial process. For $i$ from 1 to $n$, let $X_i$ be the expected number of successes in the $i$-th trial. Since there is only one $i$-th trial, this is the same as the probability of having a success, that is, $E(X_i) = p$. Furthermore, $X = X_1 + X_2 + \ldots + X_n$. Repeated application of property (i) of the Linearity of Expectation gives $E(X) = \sum_{i=1}^{n} E(X_i) = np$, as before.

**Indication.** The Linearity of Expectation does not depend on the independence of the trials; it is also true if $X$ and $Y$ are dependent. We illustrate this property by going back to our hat checking experiment. First, we introduce a definition. Given an event, the corresponding indicator random variable is 1 if the event happens and 0 otherwise. Thus, $E(X) = P(X = 1)$, where $X$ is the indicator random variable of the event.

In the hat checking experiment, we return $n$ hats in a random order. Let $X$ be the number of correctly returned hats. We proved that the probability of returning at least one hat correctly is $P(X \geq 1) = 1 - e^{-1} = 0.632 \ldots$ To compute the expectation from the definition, we would have to determine the probability of returning exactly $k$ hats corrects, for each $0 \leq k \leq n$. Alternatively, we can compute the expectation by decomposing the random variable, $X = X_1 + X_2 + \ldots + X_n$, where $X_i$ is the expected value that the $i$-th hat is returned correctly. Now, $X_i$ is an indicator variable with $E(X_i) = \frac{1}{n}$. Note that the $X_i$ are not independent. For example, if the first $n-1$ hats are returned correctly then so is the $n$-th hat. In spite of the dependence, we have

$$E(X) = \sum_{i=1}^{n} E(X_i) = 1.$$
In words, the expected number of correctly returned hats is one.

**Waiting for success.** Suppose we have again a Bernoulli trial process, but this time we end it the first time we hit a success. Defining $X$ equal to the index of the first success, we are interested in the expected value, $E(X)$. We have $P(X = i) = (1 - p)^{i-1}p$ for each $i$. As a sanity check, we make sure that the probabilities add up to one. Indeed,

$$
\sum_{i=1}^{\infty} P(X = i) = \sum_{i=1}^{\infty} (1 - p)^{i-1}p
= p \cdot \frac{1}{1 - (1 - p)}.
$$

Using the Linearity of Expectation, we get a similar sum for the expected number of trials. First, we note that for $0 \leq x < 1$, we have $\sum_{j=0}^{\infty} jx^j = \frac{x}{(1-x)^2}$. There are many ways to derive this equation, for example, by index transformation. Alternatively, you can prove

$$
\sum_{j=0}^{n} jx^j = \frac{(1-x^n)x}{(1-x)^2}
$$

by induction. Since $x^n$ goes to zero as $n$ goes to infinity, the limit of the right hand side is of course $\frac{x}{(1-x)^2}$. Hence,

$$
E(X) = \sum_{i=0}^{\infty} iP(X = i)
= \frac{p}{1 - p} \cdot \frac{1 - p}{(1 - (1 - p))^2},
$$

which is equal to $\frac{1}{2}$. This result is indeed what we intuitively think it should be, namely that we need a number of trials that is one over the probability that a single trial is a success.

**Coin flipping.** Although individual events based on probability are unpredictable, we can predict patterns when we repeat the experiment many times. Today, we will look at the pattern that emerges from independent random variables, such as flipping a coin. Suppose we have a fair coin, that is, the probability of getting head is precisely one half and the same is true for getting tail. Let $X$ count the times we get head. If we flip the coin $n$ times, the probability that we get $k$ heads is

$$
P(X = k) = \binom{n}{k} / 2^n.\]

Figure 9 visualizes this distribution in the form of a histogram for $n = 10$. Recall that the distribution function maps every possible outcome to its probability, $f(k) = P(X = k)$. This makes sense when we have a discrete domain. For a continuous domain, we consider the cumulative distribution function that gives the probability of the outcome to be within a particular range, that is,

$$
f_x(a, b) = \int_a^b f(x) \, dx = P(a \leq X \leq b).
$$

**Variance.** Now that we have an idea of what a distribution function looks like, we wish to find succinct ways of describing it. First, we note that $\mu = E(X)$ is the expected value of our random variable. It is also referred to as the mean or the average of the distribution. In the example above, where $X$ is the number of heads in ten coin flips, we have $\mu = 5$. However, we would not be surprised if we had four or six heads but we might be surprised if we had zero or ten heads when we flip a coin ten times. To express how surprised we should be, we measure the spread of the distribution. Let us first determine how close we expect a random variable to be to its expectation, $E(X - E(X))$. By linearity of expectation, we have

$$
E(X - \mu) = E(X) - E(\mu) = \mu - \mu = 0.
$$

Being always zero, this measurement is not a good description of the distribution. Instead, we use the expectation of the square of the difference to the mean. Specifically, the variance of a random variable $X$, denoted as $V(X)$, is the expectation $E((X - \mu)^2)$. The standard deviation is the square root of the variance, that is, $\sigma(X) = V(X)^{1/2}$. If $X_4$ is the number of heads we see in four coin flips, then $\mu = 2$ and

$$
V(X_4) = \frac{1}{16} \left[ (-2)^2 + 4 \cdot (-1)^2 + 4 \cdot 1^2 + 2^2 \right].
$$

Figure 9: The histogram shows the probability of getting $0 \leq i \leq 10$ heads when flipping a coin ten times.
which is equal to 1. For comparison, let $X_1$ be the number of heads that we see in one coin flip. Then $\mu = \frac{1}{2}$ and

$$V(X_1) = \frac{1}{2} \left[ (0 - \frac{1}{2})^2 + (1 - \frac{1}{2})^2 \right],$$

which is equal to one quarter. Here, we notice that the variance of four flips is the sum of the variances for four individual flips. However, this property does not hold in general.

**Additivity of Variance.** If $X$ and $Y$ are independent random variables then $V(X + Y) = V(X) + V(Y)$.

We omit the proof, not because it is difficult but because it is tedious.

**Normal distribution.** If we continue to increase the number of coin flips, then the distribution function approaches the normal distribution,

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$ 

This is the limit of the distribution as the number of coin flips approaches infinity. For a large number of trials, the normal distribution can be used to approximate the probability of the sum being between $a$ and $b$ standard deviations from the expected value.

**Standard Limit Theorem.** The probability of the number of heads being between $a\sigma$ and $b\sigma$ from the mean goes to

$$\frac{1}{\sqrt{2\pi}} \int_{x=a}^{b} e^{-\frac{x^2}{2}} \, dx$$

as the number of flips goes to infinity.

For example, if we have 100 coin flips, then $\mu = 50$, $V(X) = 25$, and $\sigma = 5$. It follows that the probability of having between 45 and 55 heads is about 0.68.