

# Chapter 4

## Expectation

שקיעה ורודה על סף הרחוב  
ורחוב כמנהרה של תכלת  
מי שיגיע עד הסוף  
ירצה לבכות מרב תוחלת

### 4.1 Basic definitions

*Definition 4.1* Let  $X$  be a real-valued random variable over a discrete probability space  $(\Omega, \mathcal{F}, P)$ . We denote the point probability by  $p(\omega) = P(\{\omega\})$ . The expectation or expected value (תוחלת) of  $X$  is a real number denoted by  $\mathbb{E}[X]$ , and defined by

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

It is an average over  $X(\omega)$ , weighted by the probability  $p(\omega)$ .

*Comment:* Mean is synonymous to expected value.

*Comment:* The expected value is only defined for random variables for which the sum converges *absolutely*. Conditional convergence is meaningless because the limit depends on the order of the summands, and the sample space does not come with any canonical order. In the context of measure theory, the expectation is the *integral* of  $X$  over the measure space  $(\Omega, \mathcal{F}, P)$ .

*Comment:* The notion of expected value of a random variable is intimately related to the statistical notion of an *average*, but it is a different notion.

The expected value of  $X$  can be rewritten in terms of the distribution of  $X$ : using the fact that

$$\Omega = \bigcup_{x \in S_X} X^{-1}(\{x\}),$$

we have

$$\begin{aligned} \mathbb{E}[X] &= \sum_{\omega \in \Omega} X(\omega) p(\omega) \\ &= \sum_{x \in S_X} \left( \sum_{\omega \in X^{-1}(\{x\})} X(\omega) p(\omega) \right) \\ &= \sum_{x \in S_X} x \sum_{\omega \in X^{-1}(\{x\})} p(\omega) \\ &= \sum_{x \in S_X} x p_X(x), \end{aligned}$$

where in the passage to the third line we used the fact that  $\omega \in X^{-1}(\{x\})$  implies that  $X(\omega) = x$ , and in the passage to the fourth line we used the definition of the point distribution  $p_X$ .

Thus,  $\mathbb{E}[X]$  is the average over all values that  $X$  may assume weighted by its point distribution, or the expected value of the identity function,  $X(x) = x$ , with respect to the probability space  $(S_X, \mathcal{F}_X, P_X)$ . This average can be calculated for random variables with discrete range, even when the probability space is not discrete. Thus, we have a second definition of expectation which extends the first one:

*Definition 4.2* Let  $X$  be a real-valued random variable with discrete range  $S_X$  and point distribution  $p_X$ . The expectation of  $X$ , denoted by  $\mathbb{E}[X]$ , is defined by

$$\mathbb{E}[X] = \sum_{x \in S_X} x p_X(x).$$

It is an average over  $S$ , weighted by the probability  $p_X(x)$ .

*Example:* Let  $X$  be the outcome of a tossed die, what is the expected value of  $X$ ?

In this case  $S = \{1, \dots, 6\}$  and  $p_X(k) = \frac{1}{6}$  for all  $k$ , thus

$$\mathbb{E}[X] = \sum_{k=1}^6 k p_X(k) = \frac{21}{6}.$$

▲ ▲ ▲

*Example:* The expected value of  $X$ , which is a Bernoulli variable with  $p_X(1) = p$ , is

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

▲ ▲ ▲

*Example:* The expected value of  $X \sim \mathcal{B}(n, p)$  is given by

$$\begin{aligned} \mathbb{E}[X] &= \sum_{k=0}^n k \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=1}^n \frac{n!}{(n-k)!(k-1)!} p^k (1-p)^{n-k} \\ &= \sum_{j=0}^{n-1} \frac{n!}{(n-j-1)!k!} p^{j+1} (1-p)^{n-j-1} \\ &= np \sum_{j=0}^{n-1} \frac{(n-1)!}{(n-1-j)!k!} p^j (1-p)^{n-1-j} \\ &= np(p + 1 - p)^{n-1} = np, \end{aligned}$$

where in the passage to the third line we changed variables,  $j = k - 1$ , and the passage to the last line follows from the binomial formula. Thus, in a binomial variable, the “weighted average” of the outcome is the probability of success in a single experiment times the number of experiments. ▲ ▲ ▲

*Example:* The expected value of  $X \sim \mathcal{Poi}(\lambda)$  is given by

$$\mathbb{E}[X] = \sum_{k=0}^{\infty} k e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} = e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^{j+1}}{j!} = \lambda.$$

Thus isn't at all surprising if we remember that in a certain sense,

$$\mathcal{Poi}(\lambda) = \lim_{n \rightarrow \infty} \mathcal{B}(n, \lambda/n).$$

▲ ▲ ▲

*Example:* What is the expected value of  $X \sim \mathcal{Geo}(p)$ ?

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} k q^{k-1} p = p \frac{d}{dq} \sum_{k=1}^{\infty} q^k = p \frac{d}{dq} \left( \frac{q}{1-q} \right) = \frac{p}{(1-q)^2} = \frac{1}{p}.$$


An alternative derivation is


$$\mathbb{E}[X] = p + \sum_{k=2}^{\infty} kq^{k-1}p = p + \sum_{k=1}^{\infty} (k+1)q^k p = p + \sum_{k=1}^{\infty} q^k p + q\mathbb{E}[X],$$

i.e.,

$$p\mathbb{E}[X] = p + \frac{pq}{1-q} = 1.$$

▲ ▲ ▲

 **Exercise 4.1** What is the expected value of the number of times one has to toss a die until getting a “3”? What is the expected value of the number of times one has to toss two dice until getting either (6, 5) or (5, 6)?

 **Exercise 4.2** Let  $X$  be a random variable assuming integer values and having a point distribution of the form  $p_X(k) = a/k^2$ , where  $a$  is a constant. What is  $a$ ? What is the expected value of  $X$ ?

**Theorem 4.1 (Expectation is monotone)** Let  $X$  and  $Y$  be random variables satisfying  $X \geq Y$  (that is,  $X(\omega) \geq Y(\omega)$  for all  $\omega \in \Omega$ ). Then,  $\mathbb{E}[X] \geq \mathbb{E}[Y]$ .

*Proof:* For all  $\omega \in \Omega$  we have  $X(\omega)p(\omega) \geq Y(\omega)p(\omega)$ . Adding up, we get

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega)p(\omega) \geq \sum_{\omega \in \Omega} Y(\omega)p(\omega) = \mathbb{E}[Y]$$

■

**Meaning of the expected value** Suppose we repeat the same experiment many times and thus obtain a sequence  $(X_k)$  of random variables that are mutually independent and have the same distribution. Consider then the *statistical average*

$$Y = \frac{1}{n} \sum_{k=1}^n X_k = \sum_{a \in S} a \frac{\text{number of times the outcome was } a}{n}.$$

As  $n$  goes to infinity, this ratio tends to  $P(X = a) = p_X(a)$ , hence  $Y$ , which is a random variable, tends to the non-random number  $\mathbb{E}[X]$ . This heuristic argument lacks rigor (e.g., does it hold when  $S$  is an infinite set?), but should give some insight into the meaning of the expected value. Note that like any average the expected value is not the value that is the most expected!

## 4.2 The expected value of a function of a random variable

*Example:* Consider a random variable  $X$  assuming the values  $\{0, 1, 2\}$  and having a distribution

$x$	0	1	2
$p_X(x)$	1/2	1/3	1/6

What is the expected value of the random variable  $Y = X^2$ ?

By definition, we need to find the distribution  $p_Y$  of  $Y$ , and then

$$\mathbb{E}[X^2] = \mathbb{E}[Y] = \sum_y y p_Y(y).$$

The distribution of  $Y$  is readily inferred from the distribution of  $X$ ,

$y$	0	1	4
$p_Y(y)$	1/2	1/3	1/6

thus

$$\mathbb{E}[Y] = \frac{1}{2} \cdot 0 + \frac{1}{3} \cdot 1 + \frac{1}{6} \cdot 4 = 1.$$

Note then that the arithmetic operation we do is equivalent to

$$\mathbb{E}[X^2] = \sum_x x^2 p_X(x).$$

The question is whether it is generally true that for any function  $g : \mathbb{R} \rightarrow \mathbb{R}$ ,

$$E[g \circ X] = \sum_x g(x) p_X(x).$$

While this may seem intuitive, note that by definition,

$$E[g \circ X] = \sum_y y p_{g \circ X}(y),$$

which seems a totally different expression. ▲ ▲ ▲

*Theorem 4.2 (The unconscious statistician (הסטטיסטיקאי הלא מודע))* Let  $X$  be a discrete random variable with range  $S_X$  and point distribution  $p_X$ . Then, for any real valued function  $g$ ,

$$\mathbb{E}[g \circ X] = \sum_{x \in S_X} g(x) p_X(x),$$

provided that the right-hand side is finite.

*Proof:* Let  $Y = g \circ X$  and set  $S_Y = g(S_X)$  be the range set of  $Y$ . We need to calculate  $\mathbb{E}[Y]$ , therefore we need to express the point distribution of  $Y$ . Let  $y \in S_Y$ , then

$$p_Y(y) = P(Y = y) = P(g \circ X = y) = P(X \in g^{-1}(\{y\})) = P_X(g^{-1}(\{y\})).$$

Thus,

$$p_Y(y) = \sum_{x \in g^{-1}(\{y\})} p_X(x).$$

The expected value of  $Y$  is


$$\begin{aligned} \mathbb{E}[Y] &= \sum_{y \in S_Y} y p_Y(y) \\ &= \sum_{y \in S_Y} y \sum_{x \in g^{-1}(\{y\})} p_X(x) \\ &= \sum_{y \in S_Y} \sum_{x \in g^{-1}(y)} y p_X(x) \\ &= \sum_{y \in S_Y} \underbrace{\sum_{x \in g^{-1}(y)} g(x) p_X(x)}_{\sum_{x \in S_X}}. \end{aligned}$$

■

*Comment:* In a sense, this theorem is trivial. Had we followed the original definition of the expected value, we would have gotten,

$$\begin{aligned}\mathbb{E}[g \circ X] &= \sum_{\omega \in \Omega} g(X(\omega)) p(\omega) \\ &= \sum_{x \in S_X} \sum_{\omega \in X^{-1}(x)} g(X(\omega)) p(\omega) \\ &= \sum_{x \in S_X} g(x) \sum_{\omega \in X^{-1}(x)} p(\omega) \\ &= \sum_{x \in S_X} g(x) p_X(x),\end{aligned}$$

except that this argument does not hold as is for a non-countable probability space.

 *Exercise 4.3* Let  $X$  be a random variable and  $f, g$  be two real valued functions. Prove that

$$\mathbb{E}[f(X)g(X)] \leq \left(\mathbb{E}[f^2(X)]\right)^{1/2} \left(\mathbb{E}[g^2(X)]\right)^{1/2}.$$

Hint: use the Cauchy inequality.

*Example:* The soccer club of Alufim FC plans to sell jerseys carrying the name of their star. They must place their order at the beginning of the year. For every sold jersey they gain  $b$  NIS, but for every jersey that remains unsold they lose  $\ell$  NIS. Suppose that the demand is a random variable with point distribution  $p(j)$ ,  $j = 0, 1, \dots$ . How many jerseys do they need to order to maximize their expected profit?

Let  $a$  be the number of jerseys ordered by the club, and  $X$  be the random demand. The net profit is then

$$g \circ X = \begin{cases} Xb - (a - X)\ell & X = 0, 1, \dots, a \\ ab & X > a \end{cases}$$

The expected gain is obtained using the law of the unconscious statistician,

$$\begin{aligned}
 \mathbb{E}[g \circ X] &= \sum_{j=1}^a [jb - (a-j)\ell] p(j) + \sum_{j=a+1}^{\infty} abp(j) \\
 &= -a\ell \sum_{j=0}^a p(j) + ab \sum_{j=a+1}^{\infty} p(j) + (b+\ell) \sum_{j=0}^a jp(j) \\
 &= ab \sum_{j=0}^{\infty} p(j) + (b+\ell) \sum_{j=0}^a (j-a)p(j) \\
 &= ab + (b+\ell) \sum_{j=0}^a (j-a)p(j) =: G(a).
 \end{aligned}$$

We need to maximize this expression with respect to  $a$ . The simplest way to do it is to check what happens when we go from  $a$  to  $a+1$ :

$$G(a+1) = G(a) + b - (b+\ell) \sum_{j=0}^a p(j).$$

That is, it is profitable to increase  $a$  as long as

$$P(X \leq a) = \sum_{j=0}^a p(j) < \frac{b}{b+\ell}.$$

▲ ▲ ▲

*Comment:* Consider a probability space  $(\Omega, \mathcal{F}, P)$ , and let  $a \in \mathbb{R}$  be a constant. We may consider  $a$  to be a constant random variable  $X(\omega) = a$ . Then,

$$p_X(a) = P(\{\omega : X(\omega) = a\}) = P(\Omega) = 1,$$


from which follows that

$$\mathbb{E}[a] = a.$$

The calculation of the expected value of a function of a random variable is easily generalized to multiple random variables. Consider a probability space  $(\Omega, \mathcal{F}, P)$  on which two random variables  $X, Y$  are defined, and let  $g : \mathbb{R}^2 \rightarrow \mathbb{R}$ . The theorem of the unconscious statistician generalizes into

$$\mathbb{E}[g(X, Y)] = \sum_{x,y} g(x, y) p_{X,Y}(x, y).$$



 **Exercise 4.4** Prove it.

**Corollary 4.1** The expectation is a linear functional in the vector space of random variables: if  $X, Y$  are random variables over a probability space  $(\Omega, \mathcal{F}, P)$  and  $a, b \in \mathbb{R}$ , then

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y].$$

*Proof:* By the theorem of the unconscious statistician,

$$\begin{aligned} \mathbb{E}[aX + bY] &= \sum_{x,y} (ax + by) p_{X,Y}(x, y) \\ &= a \sum_x x \underbrace{\sum_y p_{X,Y}(x, y)}_{p_X(x)} + b \sum_y y \underbrace{\sum_x p_{X,Y}(x, y)}_{p_Y(y)} \\ &= a\mathbb{E}[X] + b\mathbb{E}[Y]. \end{aligned}$$

■

This simple fact will be used extensively later on.

## 4.3 Moments

**Definition 4.3** Let  $X$  be a random variable over a probability space. The  $n$ -th moment (מומנט) of  $X$  is defined by

$$M_n[X] = \mathbb{E}[X^n].$$

If we denote the expected value of  $X$  by  $\mu$ , then the  $n$ -th central moment (מומנט מרכזי) of  $X$  is defined by

$$C_n[X] = \mathbb{E}[(X - \mu)^n].$$

The second central moment of a random variable is called its variance, and it is denoted by

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

The square root of the variance is called the standard deviation (סטטיית תקן), and is denoted by

$$\sigma_X = \sqrt{\text{Var}[X]}.$$

*Comments:*

1. All these definitions hold provided that the expected values exist.
2. Note that

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

where we used the linearity of the expectation.

3. The standard deviation is a measure of the (absolute) distance of the random variable from its expected value. It provides a measure of how “spread” the distribution of  $X$  is.

*Proposition 4.1* If  $\text{Var}[X] = 0$ , then  $X(\omega) = \mathbb{E}[X]$  with probability one.

*Proof:* Let  $\mu = \mathbb{E}[X]$ . By definition,

$$\text{Var}[X] = \sum_x (x - \mu)^2 p_X(x).$$

This is a sum of non-negative terms. It can only be zero if  $p_X(\mu) = 1$  and  $p_X(x) = 0$  for all  $x \neq \mu$ . ■

*Example:* The second moment of a random variable  $X \sim \mathcal{B}(n, p)$  is calculated as follows:

$$\begin{aligned} \mathbb{E}[X(X-1)] &= \sum_{k=0}^n k(k-1) \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=2}^n \frac{n!}{(n-k)!(k-2)!} p^k (1-p)^{n-k} \\ &= \sum_{j=0}^{n-2} \frac{n!}{(n-j-2)!k!} p^{j+2} (1-p)^{n-j-2} \\ &= n(n-1)p^2, \end{aligned}$$

where in the passage to third line we change the summation variable  $j = k - 2$ . Therefore,


$$\mathbb{E}[X^2] = n(n-1)p^2 + \mathbb{E}[X] = n(n-1)p^2 + np.$$

The variance of  $X$  is

$$\text{Var}[X] = n(n-1)p^2 + np - (np)^2 = np(1-p).$$

▲ ▲ ▲

*Example:* What is the variance of a Poisson variable  $X \sim \text{Poi}(\lambda)$ ? Recalling that a Poisson variable is the limit of a binomial variable with  $n \rightarrow \infty$ ,  $p \rightarrow 0$ , and  $np = \lambda$ , we deduce that  $\text{Var}[X] = \lambda$ . ▲ ▲ ▲

 *Exercise 4.5* Calculate the variance of a Poisson variable directly, without using the limit of a binomial variable.

*Proposition 4.2* For any random variable,

$$\text{Var}[aX + b] = a^2 \text{Var}[X].$$

*Proof:*

$$\text{Var}[aX + b] = \mathbb{E}[(aX + b - \mathbb{E}[aX + b])^2] = \mathbb{E}[a^2(X - \mathbb{E}[X])^2] = a^2 \text{Var}[X].$$

■

*Definition 4.4* Let  $X, Y$  be a pair of random variables. Their covariance (שונות משותפת) is defined by


$$\text{Cov}[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

Two random variables whose covariance vanishes are said to be uncorrelated (חסרי קורלציה). The correlation coefficient (מקדם קורלציה) of a pair of random variables is defined by

$$\rho[X, Y] = \frac{\text{Cov}[X, Y]}{\sigma_X \sigma_Y}.$$

The covariance of two variables is a measure of their tendency to be larger than their expected value together. A negative covariance means that when one of the variables is larger than its mean, the other is more likely to be less than its mean. Note that

$$\text{Cov}[X, Y] = \mathbb{E}[XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y] = \mathbb{E}[XY] - \mathbb{E}[X] \mathbb{E}[Y].$$

 **Exercise 4.6** Prove that the correlation coefficient of a pair of random variables assumes values between  $-1$  and  $1$  (Hint: use the Cauchy-Schwarz inequality).

**Proposition 4.3** If  $X, Y$  are independent random variables, and  $g, h$  are real valued functions, then

$$\mathbb{E}[(g \circ X)(h \circ Y)] = \mathbb{E}[g \circ X] \mathbb{E}[h \circ Y].$$

*Proof:* One only needs to apply the law of the unconscious statistician and use the fact that the joint distribution is the product of the marginal distributions,

$$\begin{aligned} \mathbb{E}[(g \circ X)(h \circ Y)] &= \sum_{x,y} g(x)h(y)p_X(x)p_Y(y) \\ &= \left( \sum_x g(x)p_X(x) \right) \left( \sum_y h(y)p_Y(y) \right) \\ &= \mathbb{E}[g \circ X] \mathbb{E}[h \circ Y]. \end{aligned}$$

■

**Corollary 4.2** If  $X, Y$  are independent then they are uncorrelated.

*Proof:* Obvious. ■

Is the opposite statement true? Are uncorrelated random variables necessarily independent? Consider the following joint distribution:

$X/Y$	$-1$	$0$	$1$
$0$	$1/3$	$0$	$1/3$
$1$	$0$	$1/3$	$0$

$X$  and  $Y$  are not independent, because, for example, knowing that  $X = 1$  implies that  $Y = 0$ . On the other hand,


$$\text{Cov}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0 - \frac{1}{3} \cdot 0 = 0.$$

That is, zero correlation does not imply independence.

*Proposition 4.4* For any two random variables  $X, Y$ ,

$$\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] + 2 \text{Cov}[X, Y].$$

*Proof:* Just do it! ■

 *Exercise 4.7* Show that for any collection of random variables  $X_1, \dots, X_n$ ,

$$\text{Var}\left[\sum_{k=1}^n X_k\right] = \sum_{k=1}^n \text{Var}[X_k] + 2 \sum_{i < j} \text{Cov}[X_i, X_j].$$

## 4.4 Using the linearity of the expectation

In this section we will examine a number of examples that exploit the additive property of the expectation.

*Example:* Recall that we calculated the expected value of a binomial variable,  $X \sim \mathcal{B}(n, p)$ , and that we obtained  $\mathbb{E}[X] = np$ . There is an easy way to obtain this result. A binomial variable can be represented as a sum of independent Bernoulli variables,

$$X = \sum_{k=1}^n X_k, \quad X_k \text{'s Bernoulli with success probability } p.$$

By the linearity of the expectation,

$$\mathbb{E}[X] = \sum_{k=1}^n \mathbb{E}[X_k] = np.$$

The variance can be obtained by the same method. We have

$$\text{Var}[X] = n \text{Var}[X_1],$$

and it remains to verify that

$$\text{Var}[X_1] = p(1-p)^2 + (1-p)(0-p)^2 = (1-p)(p(1-p) + p^2) = p(1-p).$$

Note that the calculation of the expected value does not use the independence property, whereas the calculation of the variance does. ▲ ▲ ▲

*Example:* A hundred dice are tossed. What is the expected value of their sum  $X$ ? Let  $X_k$  be the outcome of the  $k$ -th die. Since  $\mathbb{E}[X_k] = 21/6 = 7/2$ , we have by linearity,  $\mathbb{E}[X] = 100 \times 7/2 = 350$ . ▲ ▲ ▲

*Example:* Consider again the problem of the inattentive secretary who puts  $n$  letters randomly into  $n$  envelopes. What is the expected number of letters that reach their destination?

Define for  $k = 1, \dots, n$  the Bernoulli variables,

$$X_k = \begin{cases} 1 & \text{the } k\text{-th letter reached its destination} \\ 0 & \text{otherwise.} \end{cases}$$

Clearly,  $p_{X_k}(1) = 1/n$ . If  $X$  is the number of letters that reached their destination, then  $X = \sum_{k=1}^n X_k$ , and by linearity,

$$\mathbb{E}[X] = n \times \frac{1}{n} = 1.$$


We then proceed to calculate the variance. We've already seen that for a Bernoulli variable with parameter  $p$ , the variance equals  $p(1-p)$ . In this case, the  $X_k$  are not independent therefore we need to calculate their covariance. The variable  $X_1 X_2$  is also a Bernoulli variable, with parameter  $1/n(n-1)$ , so that

$$\text{Cov}[X_1, X_2] = \frac{1}{n(n-1)} - \frac{1}{n^2}.$$

Putting things together,

$$\begin{aligned} \text{Var}[X] &= n \times \frac{1}{n} \left(1 - \frac{1}{n}\right) + 2 \binom{n}{2} \times \left(\frac{1}{n(n-1)} - \frac{1}{n^2}\right) \\ &= \left(1 - \frac{1}{n}\right) + n(n-1) \left(\frac{1}{n(n-1)} - \frac{1}{n^2}\right) = 1. \end{aligned}$$

Should we be surprised? Recall that  $X$  tends as  $n \rightarrow \infty$  to a Poisson variable with parameter  $\lambda = 1$ , so that we expect that in this limit  $\mathbb{E}[X] = \text{Var}[X] = 1$ . It turns out that this result holds exactly for every finite  $n$ . ▲ ▲ ▲

 *Exercise 4.8* In an urn are  $N$  white balls and  $M$  black balls.  $n$  balls are drawn randomly. What is the expected value of the number of white balls that were drawn? (Solve this problem by using the linearity of the expectation.)

*Example:* Consider a randomized deck of  $2n$  cards, two of type “1”, two of type “2”, and so on.  $m$  cards are randomly drawn. What is the expected value of the number of pairs that will remain intact? (This problem was solved by Daniel Bernoulli in the context of the number of married couples remaining intact after  $m$  deaths.)

We define  $X_k$  to be a Bernoulli variable taking the value 1 if the  $k$ -th couple remains intact. We have

$$\mathbb{E}[X_k] = p_{X_k}(1) = \frac{\binom{2n-2}{m}}{\binom{2n}{m}} = \frac{(2n-m)(2n-m-1)}{2n(2n-1)}.$$

The desired result is  $n$  times this number. ▲ ▲ ▲

*Example:* Recall the coupon collector: there are  $n$  different coupons, and each turn there is an equal probability to obtain any coupon. What is the expected value of the number of coupons that need to be collected before obtaining a complete set?

First, make sure you remember what the is probability space. For  $k = 0, \dots, n$ , let

$$T_k(\omega) = \text{time until } k \text{ different coupons collected.}$$

Then,  $T_n$  is the total number of coupons that need to be gathered and

$$X_k(\omega) = T_{k+1}(\omega) - T_k(\omega)$$


is the number of coupons that need to be gathered from the moment that we had  $k$  different coupons to the moment we have  $k + 1$  different coupons. Clearly,

$$T_n = \sum_{k=0}^{n-1} X_k.$$

Now, suppose we have  $k$  different coupons. Every new coupon can be viewed as a Bernoulli experiment with success probability  $(n - k)/n$ . Thus,  $X_k$  is a geometric variable with parameter  $(n - k)/n$ , and  $\mathbb{E}[X_k] = n/(n - k)$ . Summing up,

$$\mathbb{E}[T_n] = \sum_{k=0}^{n-1} \frac{n}{n-k} = n \sum_{k=1}^n \frac{1}{k} \approx n \log n.$$

▲ ▲ ▲

 *Exercise 4.9* Let  $X_1, \dots, X_n$  be a sequence of independent random variables that have the same distribution. We denote  $\mathbb{E}[X_1] = \mu$  and  $\text{Var}[X_1] = \sigma^2$ . Find the expected value and the variance of the empirical mean

$$S_n = \frac{1}{n} \sum_{k=1}^n X_k.$$

We conclude this section with a remark about infinite sums. First a simple lemma:

*Lemma 4.1* Let  $X$  be a random variable. If  $X(\omega) \geq a$  (with probability one) then  $\mathbb{E}[X] \geq a$ . Also,

$$\mathbb{E}[|X|] \geq |\mathbb{E}[X]|.$$

*Proof:* The first result follows from the definition of the expectation. The second result follows from the inequality

$$\left| \sum_x x p_X(x) \right| \leq \sum_x |x| p_X(x).$$

■

*Theorem 4.3* Let  $(X_n)$  be an infinite sequence of random variables such that

$$\sum_{n=1}^{\infty} \mathbb{E}[|X_n|] < \infty.$$

Then,

$$\mathbb{E} \left[ \sum_{n=1}^{\infty} X_n \right] = \sum_{n=1}^{\infty} \mathbb{E}[X_n].$$



*Proof:* TO BE COMPLETED. ■

The following is an application of the above theorem. Let  $X$  be a random variable assuming positive integer values and having a finite expectation. Define for every natural  $i$ ,

$$X_i(\omega) = \begin{cases} 1 & i \leq X(\omega) \\ 0 & \text{otherwise} \end{cases}$$

Then,

$$\sum_{i=1}^{\infty} X_i(\omega) = \sum_{i \leq X(\omega)} 1 = X(\omega).$$

Now,

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} \mathbb{E}[X_i] = \sum_{i=1}^{\infty} P(\{\omega : X(\omega) \geq i\}).$$

## 4.5 Conditional expectation

*Definition 4.5* Let  $X, Y$  be random variables over a probability space  $(\Omega, \mathcal{F}, P)$ . The conditional expectation of  $X$  given that  $Y = y$  is defined by

$$\mathbb{E}[X | Y = y] = \sum_x x p_{X|Y}(x | y).$$

Note that this definition makes sense because  $p_{X|Y}(\cdot | y)$  is a point distribution on  $S_X$ .

*Example:* Let  $X, Y \sim \mathcal{B}(n, p)$  be independent. What is the conditional expectation of  $X$  given that  $X + Y = m$ ?

To answer this question we need to calculate the conditional distribution  $p_{X|X+Y}$ . Now,

$$p_{X|X+Y}(k | m) = \frac{P(X = k, X + Y = m)}{P(X + Y = m)} = \frac{P(X = k, Y = m - k)}{P(X + Y = m)},$$

with  $k \leq m, n$ . We know what the numerator is. For the denominator, we realize that the sum of two binomial variables with parameters  $(n, p)$  is a binomial variable with parameters  $(2n, p)$  (think of two independent sequences of Bernoulli trials added up). Thus,

$$p_{X|X+Y}(k | m) = \frac{\binom{n}{k} p^k (1-p)^{n-k} \binom{n}{m-k} p^{m-k} (1-p)^{n-m+k}}{\binom{2n}{m} p^m (1-p)^{2n-m}} = \frac{\binom{n}{k} \binom{n}{m-k}}{\binom{2n}{m}}.$$

The desired result is

$$\mathbb{E}[X \mid X + Y = m] = \sum_{k=0}^{\min(m,n)} k \frac{\binom{n}{k} \binom{n}{m-k}}{\binom{2n}{m}}.$$

It is not clear how to simplify this expression. A useful trick is to observe that  $p_{X|X+Y}(k \mid m)$  with  $m$  fixed is the probability of obtaining  $k$  white balls when one draws  $m$  balls from an urn containing  $n$  white balls and  $n$  black balls. Since every ball is white with probability  $1/2$ , by the linearity of the expectation, the expected number of white balls is  $m/2$ .

Now that we know the result, we may see that we could have reached it much more easily. By symmetry,

$$\mathbb{E}[X \mid X + Y = m] = \mathbb{E}[Y \mid X + Y = m],$$

hence, by the linearity of the expectation,

$$\mathbb{E}[X \mid X + Y = m] = \frac{1}{2} \mathbb{E}[X + Y \mid X + Y = m] = \frac{m}{2}.$$

In particular, this result holds whatever the distribution of  $X, Y$  is (as long as it is the same). ▲ ▲ ▲

We now refine our definition of the conditional expectation:

*Definition 4.6* Let  $X, Y$  be random variables over a probability space  $(\Omega, \mathcal{F}, P)$ . The conditional expectation of  $X$  given  $Y$  is a random variable  $Z$ , which is a composite function of  $Y$ , i.e.,  $Z(\omega) = \varphi(Y(\omega))$ , and

$$\varphi(y) = \mathbb{E}[X \mid Y = y].$$

Another way to write it is:

$$\mathbb{E}[X \mid Y](\omega) = \mathbb{E}[X \mid Y = y]_{y=Y(\omega)}.$$

That is, having performed the experiment, we are given only  $Y(\omega)$ , and the random variable  $\mathbb{E}[X \mid Y](\omega)$  is the expected value of  $X$  now that we know  $Y(\omega)$ .

*Example:* Returning to the above example,

$$\mathbb{E}[X \mid X + Y = m] = \frac{m}{2},$$

hence

$$\mathbb{E}[X | X + Y](\omega) = \frac{X(\omega) + Y(\omega)}{2}.$$

▲ ▲ ▲

*Proposition 4.5* For every two random variables  $X, Y$ ,

$$\mathbb{E}[\mathbb{E}[X | Y]] = \mathbb{E}[X].$$

*Proof:* What does the proposition say? That

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

Since  $\mathbb{E}[X | Y](\omega)$  is a composite function of  $Y(\omega)$  we can use the law of the unconscious statistician to rewrite it as

$$\sum_y \mathbb{E}[X | Y = y] p_Y(y) = \sum_x x p_X(x).$$

Indeed,

$$\begin{aligned} \sum_y \mathbb{E}[X | Y = y] p_Y(y) &= \sum_y \sum_x x p_{X|Y}(x | y) p_Y(y) \\ &= \sum_y \sum_x x p_{X,Y}(x, y) = \mathbb{E}[X]. \end{aligned}$$

■

This simple proposition is quite useful. It states that the expected value of  $X$  can be computed by averaging over its expectation conditioned over another variable.

*Example:* A miner is inside a mine, and doesn't know which of three possible tunnels will lead him out. If he takes tunnel A he will be out within 3 hours. If he takes tunnel B he will be back to the same spot after 5 hours. If he takes tunnel C he will be back to the same spot after 7 hours. He chooses the tunnel at random with equal probability for each tunnel. If he happens to return to the same spot, the poor thing is totally disoriented, and has to redraw his choice again with equal probabilities. What is the expected time until he finds the exit?

The sample space consists of infinite sequences of "BCACCBA...", with the standard probability of independent repeated trials. Let  $X(\omega)$  be the exit time and  $Y(\omega)$  be the label of the first door he chooses. By the above proposition,

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}[\mathbb{E}[X | Y]] \\ &= \mathbb{E}[X | Y = A] p_Y(A) + \mathbb{E}[X | Y = B] p_Y(B) + \mathbb{E}[X | Y = C] p_Y(C) \\ &= \frac{1}{3} (3 + \mathbb{E}[X | Y = B] + \mathbb{E}[X | Y = C]).\end{aligned}$$

What is  $\mathbb{E}[X | Y = B]$ ? If the miner chose tunnel B, then he wandered for 5 hours, and then faced again the original problem, independently of his first choice. Thus,

$$\mathbb{E}[X | Y = B] = 5 + \mathbb{E}[X] \quad \text{and similarly} \quad \mathbb{E}[X | Y = C] = 7 + \mathbb{E}[X].$$

Substituting, we get

$$\mathbb{E}[X] = 1 + \frac{1}{3} (5 + \mathbb{E}[X]) + \frac{1}{3} (7 + \mathbb{E}[X]).$$

This equation is easily solved,  $\mathbb{E}[X] = 15$ . ▲ ▲ ▲

*Example:* Consider a sequence of independent Bernoulli trials with success probability  $p$ . What is the expected number of trials until one obtains two 1's in a row?

Let  $X(\omega)$  be the number of trials until two 1's in a row, and let  $Y_j(\omega)$  be the outcome of the  $j$ -th trial. We start by writing

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X | Y_1]] = p \mathbb{E}[X | Y_1 = 1] + (1 - p) \mathbb{E}[X | Y_1 = 0].$$

By the same argument as above,

$$\mathbb{E}[X | Y_1 = 0] = 1 + \mathbb{E}[X].$$

Next, we use a simple generalization of the conditioning method,

$$\mathbb{E}[X | Y_1 = 1] = p \mathbb{E}[X | Y_1 = 1, Y_2 = 1] + (1 - p) \mathbb{E}[X | Y_1 = 1, Y_2 = 0].$$

Using the fact that

$$\mathbb{E}[X | Y_1 = 1, Y_2 = 1] = 2 \quad \text{and} \quad \mathbb{E}[X | Y_1 = 1, Y_2 = 0] = 2 + \mathbb{E}[X],$$

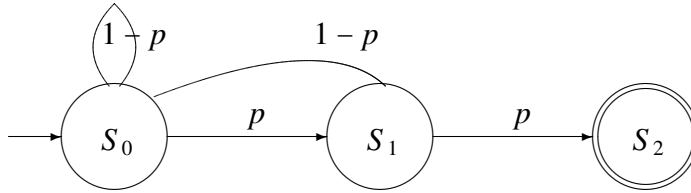
we finally obtain an implicit equation for  $\mathbb{E}[X]$ :

$$\mathbb{E}[X] = p [2p + (1 - p)(2 + \mathbb{E}[X])] + (1 - p)(1 + \mathbb{E}[X]),$$

from which we readily obtain

$$\mathbb{E}[X] = \frac{1 + p}{p^2}.$$

We can solve this same problem differently. We view the problem in terms of a three-state space: one can be in the initial state (having to produce two 1's in a row), be in state after a single 1, or be in the terminal state after two 1's in a row. We label these states  $S_0$ ,  $S_1$ , and  $S_2$ . Now every sequence of successes and failures implies a trajectory on the state space. That is, we can replace the original sample space of sequences of zero-ones by a sample space of sequences of states  $S_j$ . This defines a new compound experiment, with transition probabilities that can be represented as a graph:





Let now  $X(\omega)$  be the number of steps until reaching state  $S_2$ . The expected value of  $X$  depends on the initial state. The graph suggests the following relations,

$$\begin{aligned}\mathbb{E}[X \mid S_0] &= 1 + p\mathbb{E}[X \mid S_1] + (1 - p)\mathbb{E}[X \mid S_0] \\ \mathbb{E}[X \mid S_1] &= 1 + p\mathbb{E}[X \mid S_2] + (1 - p)\mathbb{E}[X \mid S_0] \\ \mathbb{E}[X \mid S_2] &= 0\end{aligned}$$

It is easily checked that  $\mathbb{E}[X \mid S_0] = (1 + p)/p^2$ .

▲ ▲ ▲

 **Exercise 4.10** Consider a sequence of independent Bernoulli trials with success probability  $p$ . What is the expected number of trials until one obtains three 1's in a row? four 1's in a row?

 **Exercise 4.11** A monkey types randomly on a typing machine. Each character has a probability of  $1/26$  of being each of the letters of the alphabet, independently of the other. What is the expected number of characters that the monkey will type until generating the string "ABCD"? What about the string "ABAB"?

The following paragraphs are provided for those who want to know more.

The conditional expectation  $\mathbb{E}[X | Y](\omega)$  plays a very important role in probability theory. Its formal definition, which remains valid in the general case (i.e., uncountable spaces), is somewhat more involved than that presented in this section, but we do have all the necessary background to formulate it. Recall that a random variable  $Y(\omega)$  generates a  $\sigma$ -algebra of events (a sub- $\sigma$ -algebra of  $\mathcal{F}$ ),

$$\mathcal{F} \supseteq \sigma(Y) = \{Y^{-1}(A) : A \in \mathcal{F}_Y\}.$$

Let  $\varphi$  be a real valued function defined on  $S_Y$ , and define a random variable  $Z(\omega) = \varphi(Y(\omega))$ . The  $\sigma$ -algebra generated by  $Z$  is

$$\sigma(Z) = \{Z^{-1}(B) : B \in \mathcal{F}_Z\} = \{Y^{-1}(\varphi^{-1}(B)) : B \in \mathcal{F}_Z\} \subseteq \sigma(Y).$$

That is, the  $\sigma$ -algebra generated by a function of a random variable is contained in the  $\sigma$ -algebra generated by this random variable. In fact, it can be shown that the opposite is true. If  $Y, Z$  are random variables and  $\sigma(Z) \subseteq \sigma(Y)$ , then  $Z$  can be expressed as a composite function of  $Y$ .

Recall now our definition of the conditional expectation,

$$\mathbb{E}[X | Y](\omega) = \mathbb{E}[X | Y = y]_{y=Y(\omega)} = \sum_x x p_{X|Y}(x | Y(\omega)) = \sum_x x \frac{p_{X,Y}(x, Y(\omega))}{p_Y(Y(\omega))}.$$

Let  $A \in \mathcal{F}$  be any event in  $\sigma(Y)$ , that is, there exists a  $B \in \mathcal{F}_Y$  for which  $Y^{-1}(B) = A$ . Now,

$$\begin{aligned} \sum_{\omega \in A} \mathbb{E}[X | Y](\omega) p(\omega) &= \sum_{\omega \in A} \sum_x x \frac{p_{X,Y}(x, Y(\omega))}{p_Y(Y(\omega))} p(\omega) \\ &= \sum_x x \sum_{\omega \in A} \frac{p_{X,Y}(x, Y(\omega))}{p_Y(Y(\omega))} p(\omega) \\ &= \sum_x x \sum_{y \in B} \sum_{\omega \in Y^{-1}(y)} \frac{p_{X,Y}(x, y)}{p_Y(y)} p(\omega) \\ &= \sum_x x \sum_{y \in B} \frac{p_{X,Y}(x, y)}{p_Y(y)} \sum_{\omega \in Y^{-1}(y)} p(\omega) \\ &= \sum_x x \sum_{y \in B} p_{X,Y}(x, y). \end{aligned}$$

On the other hand,

$$\begin{aligned}
 \sum_{\omega \in A} X(\omega) p(\omega) &= \sum_{y \in B} \sum_{\omega \in Y^{-1}(y)} X(\omega) p(\omega) \\
 &= \sum_x \sum_{y \in B} \sum_{\{\omega: (X(\omega), Y(\omega)) = (x, y)\}} X(\omega) p(\omega) \\
 &= \sum_x x \sum_{y \in B} \sum_{\{\omega: (X(\omega), Y(\omega)) = (x, y)\}} p(\omega) \\
 &= \sum_x x \sum_{y \in B} p_{X,Y}(x, y).
 \end{aligned}$$

That is, for every  $A \in \sigma(Y)$ ,

$$\sum_{\omega \in A} \mathbb{E}[X | Y](\omega) p(\omega) = \sum_{\omega \in A} X(\omega) p(\omega).$$

This property is in fact the standard definition of the conditional expectation:

*Definition 4.7* Let  $X, Y$  be random variables over a probability space  $(\Omega, \mathcal{F}, P)$ . The conditional expectation of  $X$  given  $Y$  is a random variable  $Z$  satisfying the following two properties: (1)  $\sigma(Z) \subseteq \sigma(Y)$ , (2) For every  $A \in \sigma(Y)$

$$\sum_{\omega \in A} Z(\omega) p(\omega) = \sum_{\omega \in A} X(\omega) p(\omega).$$

It can be proved that there exists a unique random variable satisfying these properties.

## 4.6 The moment generating function

*Definition 4.8* Let  $X$  be a discrete random variable. Its moment generating function (פונקציית יוצרת מומנטים)  $M_X(t)$  is a real-valued function defined by

$$M_X(t) = \mathbb{E}[e^{tX}] = \sum_x e^{tx} p_X(x).$$

*Example:* The moment generating function of a binomial variable  $X \sim \mathcal{B}(n, p)$  is

$$M_X(t) = \sum_{k=1}^n e^{tk} \binom{n}{k} p^k (1-p)^{n-k} = (1-p + e^t p)^n.$$


▲ ▲ ▲

Note that if  $X$  and  $Y$  are independent random variables then  $M_{X+Y}(t) = M_X(t)M_Y(t)$ . Thus, it is enough to calculate  $M_X(t)$  for  $X \sim \mathcal{B}(1, p)$ , which is straightforward.

*Example:* The moment generating function of a Poisson variable  $X \sim \text{Poi}(\lambda)$  is

$$M_X(t) = \sum_{k=0}^{\infty} e^{tk} e^{-\lambda} \frac{\lambda^k}{k!} = e^{\lambda(e^t-1)}.$$


▲ ▲ ▲

 *Exercise 4.12* Find the moment generating function of  $X \sim \text{Geo}(p)$ . Is it defined for every  $t$ ?

We now explain the origin of the name *moment generating function*. Note that

$$\begin{aligned} M_X(0) &= 1 \\ M'_X(0) &= \mathbb{E}[X] \\ M''_X(0) &= \mathbb{E}[X^2], \end{aligned}$$

and in general, the  $k$ -th derivative evaluated at zero equals to the  $k$ -th moment.

 *Exercise 4.13* Verify that we get the correct moments for the binomial, Poisson and geometric random variables

*Comment:* The moment-generating function is the *Laplace transform* of the point distribution. It has many uses. We will see one of them in the section about Hoeffding's inequality.