

# Chapter 3

## Random Variables (Discrete Case)

### 3.1 Basic Definitions

Consider a probability space  $(\Omega, \mathcal{F}, P)$  corresponding to an experiment. The points  $\omega \in \Omega$  represent all possible outcomes of the experiment. In many cases, we are not necessarily interested in the outcome  $\omega$  itself, but rather in some property (function) of it. Consider the following pedagogical example: in a coin toss, a perfectly legitimate sample space is the set of initial conditions of the toss (position, velocity and angular velocity of the toss, complemented perhaps with wind conditions). Yet, all we are interested in is a very complicated function of this sample space: whether the coin ended up with Head or Tail showing up. The following is a “preliminary” version of a definition that will be refined further below:

*Definition 3.1* Let  $(\Omega, \mathcal{F}, P)$  be a probability space. A function  $X : \Omega \rightarrow S$  (where  $S$  is a set) is called a random variable (משתנה מקרי).

*Example:* Two dice are tossed, i.e.,

$$\Omega = \{(i, j) : 1 \leq i, j \leq 6\}.$$

The random variable  $X$  is the sum, i.e.,

$$X((i, j)) = i + j.$$

Note that the set  $S$  (the range of  $X$ ) can be chosen to be  $\{2, \dots, 12\}$ . Suppose now that all our probabilistic interest is in the value of  $X$ , rather than the outcome of

the individual dice (this would be the case if we played *snakes and ladders*). In such case, it seems reasonable to construct a new probability space in which the sample space is  $S$ . Since it is a finite space, the events related to  $X$  can be taken to be all subsets of  $S$ . But now we need a probability function on  $(S, 2^S)$ , which is compatible with the experiment. If  $A \in 2^S$  (e.g., the sum was greater than 5), the probability that  $A$  has occurred is given by

$$P(\{\omega \in \Omega : X(\omega) \in A\}) = P(\{\omega \in \Omega : \omega \in X^{-1}(A)\}) = P(X^{-1}(A)).$$

That is, the probability function associated with the experiment  $(S, 2^S)$  is  $P \circ X^{-1}$ . We call it the *distribution* (התפלגות) of the random variable  $X$  and denote it by  $P_X$ .

▲ ▲ ▲

**Generalization** These notions need to be formalized and generalized. In probability theory, a space (the sample space) comes with a structure (a  $\sigma$ -algebra of events). Thus, when we consider a function from the sample space  $\Omega$  to some other space  $S$ , this other space must come with its own structure—its own  $\sigma$ -algebra of events, which we denote by  $\mathcal{F}_S$ .

The function  $X : \Omega \rightarrow S$  is not necessarily one-to-one (although it can always be made onto by restricting  $S$  to be the range of  $X$ ), therefore  $X$  is not necessarily invertible. However, as we have seen,  $X^{-1}$  is well-defined on subsets of  $S$ ,

$$X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\}, \quad \forall A \in \mathcal{F}_S.$$

There is however nothing to guarantee that for every event  $A \in \mathcal{F}_S$  the set  $X^{-1}(A) \subset \Omega$  is an event in  $\mathcal{F}$ . This is something we want to ensure otherwise it will make no sense to ask “what is the probability that  $X(\omega) \in A$ ?”.

*Definition 3.2* Let  $(\Omega, \mathcal{F})$  and  $(S, \mathcal{F}_S)$  be two measurable spaces. A function  $X : \Omega \rightarrow S$  is called a random variable if  $X^{-1}(A) \in \mathcal{F}$  for all  $A \in \mathcal{F}_S$ . In the context of measure theory it is called a measurable function (פונקציה מדידה).<sup>1</sup>

*Comment:*  $X$  is a random variable if the  $\sigma$ -algebra

$$\{X^{-1}(A) : A \in \mathcal{F}_S\}$$

is a sub- $\sigma$ -algebra of  $\mathcal{F}$ .

---

<sup>1</sup>Note the analogy with the definition of continuous functions between topological spaces, the definition of linear transformations between vector spaces, and the definition of homomorphisms between groups.

*Example:* Let  $A$  be an event in a measurable space  $(\Omega, \mathcal{F})$ . An event is not a random variable, however, we can always form from an event a binary random variable (a *Bernoulli variable*), as follows. Let  $S = \{0, 1\}$  and  $\mathcal{F}_S = 2^S$ . Then,  $I_A : \Omega \rightarrow S$  is defined by

$$I_A(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \text{otherwise} \end{cases}.$$

$I_A$  is called the *indicator function* (פונקציה מציינת) of the event  $A$ .  $I_A = 1$  means that the event  $A$  has occurred. Note that  $I_A$  is measurable, as

$$I_A^{-1}(\{1\}) = \{\omega \in \Omega : I_A(\omega) = 1\} = \{\omega \in \Omega : \omega \in A\} = A \in \mathcal{F}$$

and similarly  $I_A^{-1}(\{0\}) = A^c \in \mathcal{F}$ . ▲ ▲ ▲

So far, we completely ignored probabilities and only concentrated on the structure that the function  $X$  induces on the measurable spaces that are its domain and range. Now, we remember that a probability function  $P$  is defined on  $(\Omega, \mathcal{F})$ . We want to define the probability function that it induces on  $(S, \mathcal{F}_S)$ .

*Definition 3.3* Let  $X$  be an  $(S, \mathcal{F}_S)$ -valued random variable on a probability space  $(\Omega, \mathcal{F}, P)$ . The distribution (התפלגות)  $P_X$  of  $X$  is a function  $\mathcal{F}_S \rightarrow \mathbb{R}$  defined by

$$P_X(A) = P(X^{-1}(A)) = P(\{\omega \in \Omega : X(\omega) \in A\}).$$

*Comment:* In short-hand notation,

$$P_X(A) = P(X \in A).$$

*Proposition 3.1* The distribution  $P_X$  is a probability function on  $(S, \mathcal{F}_S)$ .

*Proof:* The range of  $P_X$  is obviously  $[0, 1]$ . Also

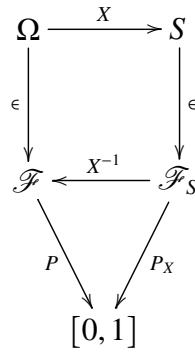
$$P_X(S) = P(X^{-1}(S)) = P(\Omega) = 1.$$

Finally, let  $(A_n) \subset \mathcal{F}_S$  be a sequence of disjoint events, then

$$\begin{aligned} P_X \left( \bigcup_{n=1}^{\infty} A_n \right) &= P \left( X^{-1} \left( \bigcup_{n=1}^{\infty} A_n \right) \right) = P \left( \bigcup_{n=1}^{\infty} X^{-1}(A_n) \right) \\ &= \sum_{n=1}^{\infty} P(X^{-1}(A_n)) = \sum_{n=1}^{\infty} P_X(A_n), \end{aligned}$$

where the first and last equalities follow from the definition of  $P_X$ , the second equality follows from the commutation of inverse functions with set-theoretic operations and the third equality follows from the fcountable additivity of probability. ■

*Comment:* The distribution is defined such that the following diagram commutes



In this chapter, we restrict our attention to random variables whose ranges  $S$  are countable spaces, and take  $\mathcal{F}_S = 2^S$ . Then the distribution is fully specified by its value for all singletons,

$$P_X(\{s\}) =: p_X(s), \quad s \in S.$$

We call the function  $p_X$  the *point distribution of the random variable  $X$*  (התפלגות נקודתית). Note the following identity,

$$p_X(s) = P_X(\{s\}) = P(X^{-1}(\{s\})) = P(\{\omega \in \Omega : X(\omega) = s\}),$$

where

$$P : \mathcal{F} \rightarrow [0, 1] \quad P_X : \mathcal{F}_S \rightarrow [0, 1] \quad p_X : S \rightarrow [0, 1]$$

are the probability, the distribution of  $X$ , and the point distribution of  $X$ , respectively. The function  $p_X$  is also called the *probability mass function* (PMF) of the random variable  $X$ .

*Example:* Three balls are extracted from an urn containing 20 balls numbered from one to twenty. What is the probability that at least one of the three has a number 17 or higher.

This question can easily be answered without random variables, but we will introduce a random variables for didactic reasons. We take the sample space to be

$$\Omega = \{(i, j, k) : 1 \leq i < j < k \leq 20\},$$

and for every  $\omega \in \Omega$ ,  $p(\omega) = 1/\binom{20}{3}$ . We define the random variable

$$X((i, j, k)) = k,$$

which returns the largest number drawn. It maps every point  $\omega \in \Omega$  into a point in the set  $S = \{3, \dots, 20\}$ . To every  $k \in S$  corresponds an event in  $\mathcal{F}$ ,

$$X^{-1}(\{k\}) = \{(i, j, k) : 1 \leq i < j < k\} \in \mathcal{F}.$$

The point distribution of  $X$  is

$$p_X(k) = P_X(\{k\}) = P(X = k) = \frac{\binom{k-1}{2}}{\binom{20}{3}}.$$

Then,

$$\begin{aligned} P_X(\{17, \dots, 20\}) &= p_X(17) + p_X(18) + p_X(19) + p_X(20) \\ &= \binom{20}{3}^{-1} \left\{ \binom{16}{2} + \binom{17}{2} + \binom{18}{2} + \binom{19}{2} \right\} \approx 0.508. \end{aligned}$$

▲ ▲ ▲

*Example:* Let  $A$  be an event in a probability space  $(\Omega, \mathcal{F}, P)$ . We have already defined the random variables  $I_A : \Omega \rightarrow \{0, 1\}$ . The distribution of  $I_A$  is determined by its value for the two singletons  $\{0\}, \{1\}$ . Now,

$$p_{I_A}(1) = P_{I_A}(\{1\}) = P(I_A^{-1}(\{1\})) = P(\{\omega : I_A(\omega) = 1\}) = P(A).$$

▲ ▲ ▲

*Example:* **The coupon collector problem.** Consider the following situation: there are  $N$  types of coupons. A coupon collector gets each time unit a coupon at

random. The probability of getting each time a specific coupon is  $1/N$ , independently of prior selections. Thus, our sample space consists of infinite sequences of coupon selections,  $\Omega = \{1, \dots, N\}^{\mathbb{N}}$ , and for every finite sub-sequence the corresponding probability space is that of equal probability. For example, setting  $\Omega_0 = \{1, \dots, N\}$ ,

$$P(\{(1, 2, 5, 17, 1, 1, 4)\} \times \Omega_0^{\mathbb{N}}) = \frac{1}{N^7}.$$

A random variable of particular interest is the number of time units  $T$  until the coupon collector has gathered at least one coupon of each sort. This random variable takes values in the set  $S = \{N, N+1, \dots\} \cup \{\infty\}$ . Our goal is to compute its point distribution  $p_T(k)$ . It is easy to see, for example, that

$$p_T(N) = P_T(\{N\}) = P(T = N) = \frac{N!}{N^N} \sim \frac{\sqrt{2\pi} N^{N+1/2} e^{-N}}{N^N} = \sqrt{2\pi N} e^{-N},$$

where we have used here *Stirling's formula*, whereby

$$\lim_{N \rightarrow \infty} \frac{\sqrt{2\pi} N^{N+1/2} e^{-N}}{N!} = 1.$$

Fix an integer  $n \geq N$ , and define the events  $A_1, A_2, \dots, A_N$  such that  $A_j$  is the event that no type- $j$  coupon is among the first  $n$  coupons. By the inclusion-exclusion principle,

$$\begin{aligned} P_T(\{k : k > n\}) &= P\left(\bigcup_{j=1}^N A_j\right) \\ &= \sum_j P(A_j) - \sum_{j < k} P(A_j \cap A_k) + \dots \end{aligned}$$

Now, by the independence of selections,  $P(A_j) = [(N-1)/N]^n$ ,  $P(A_j \cap A_k) = [(N-2)/N]^n$ , and so on, so that

$$\begin{aligned} P_T(\{k : k > n\}) &= N \left(\frac{N-1}{N}\right)^n - \binom{N}{2} \left(\frac{N-2}{N}\right)^n + \dots \\ &= \sum_{j=1}^N (-1)^{j+1} \binom{N}{j} \left(\frac{N-j}{N}\right)^n. \end{aligned}$$

Finally,

$$p_T(n) = P_T(\{k : k > n-1\}) - P_T(\{k : k > n\}).$$

Note that the coupon collector problem is in a sense dual to the birthday problem. In the latter, we are interested in when two identical birthdays are obtained for the first time. In the former, we are interested in the first time that all possible birthdays were obtained. ▲ ▲ ▲

(12 hrs)  (12 hrs)

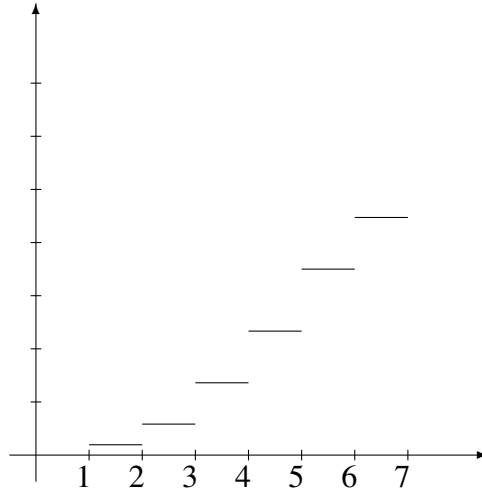
## 3.2 The cumulative distribution function

*Definition 3.4* Let  $X : \Omega \rightarrow S$  be a real-valued random variable. Its cumulative distribution function (פונקציית ההפלגות מצטברת)  $F_X$  is a real-valued function  $\mathbb{R} \rightarrow \mathbb{R}$  defined by

$$F_X(x) = P(\{\omega : X(\omega) \leq x\}) = P(X \leq x) = P_X((-\infty, x]).$$

*Comment:* Strictly speaking, for  $F_X$  to be well-defined, the pre-images of all closed rays  $X^{-1}((-\infty, x])$  must be in  $\mathcal{F}$ , and for  $P_X((-\infty, x])$  to be defined we must have  $(-\infty, x] \in \mathcal{F}_S$ . This may seem a problem if  $S$  is countable and  $\mathcal{F}_S = 2^S$ . Without getting into details that belong to measure theory, we will state that if we take  $S = \mathbb{R}$  (i.e., the range of  $X$  is larger than its image), then we can endow  $\mathbb{R}$  with a  $\sigma$ -algebra generated by all the sets of the form  $(-\infty, x]$ ; it is called the  $\sigma$ -algebra of *Borel sets*.

*Example:* Consider the experiment of tossing two dice and the random variable  $X(i, j) = i + j$ . Then,  $F_X(x)$  is of the form



▲ ▲ ▲

*Proposition 3.2* The cumulative distribution function  $F_X$  of any random variable  $X$  satisfies the following properties:

1.  $F_X$  is non-decreasing.
2.  $F_X(x)$  tends to zero when  $x \rightarrow -\infty$ .
3.  $F_X(x)$  tends to one when  $x \rightarrow \infty$ .
4.  $F_X$  is right-continuous.

*Proof:*

1. Let  $a \leq b$ , then  $(-\infty, a] \subseteq (-\infty, b]$  and since  $P_X$  is a probability function,

$$F_X(a) = P_X((-\infty, a]) \leq P_X((-\infty, b]) = F_X(b).$$

2. Let  $(x_n)$  be a sequence that converges to  $-\infty$ . Then, by the continuity of probability (for decreasing sequences)


$$\lim_n F_X(x_n) = \lim_n P_X((-\infty, x_n]) = P_X(\lim_n (-\infty, x_n]) = P_X(\emptyset) = 0.$$



3. The other limit is treated along the same lines.
4. Same for right continuity: if the sequence  $(h_n)$  converges to zero from the right, then

$$\begin{aligned}
 \lim_n F_X(x + h_n) &= \lim_n P_X((-\infty, x + h_n]) \\
 &= P_X(\lim_n (-\infty, x + h_n]) \\
 &= P_X((-\infty, x]) = F_X(x).
 \end{aligned}$$

■

 *Exercise 3.1* Explain why  $F_X$  is not necessarily left-continuous.

What is the importance of the cumulative distribution function? A distribution is a complicated object, as it has to assign a number to any set in the range of  $X$  (for the moment, let's forget that we deal with discrete variables and consider the more general case where  $S$  may be a continuous subset of  $\mathbb{R}$ ). The cumulative distribution function is a real-valued function (much simpler object) which encodes the same information. That is, the cumulative distribution function defines uniquely the distribution of any (measurable) set in  $\mathbb{R}$ . For example, the distribution of semi-open segments is

$$P_X((a, b]) = P_X((-\infty, b] \setminus (-\infty, a]) = F_X(b) - F_X(a).$$

What about open segments?

$$P_X((a, b)) = P_X(\lim_n (a, b - 1/n]) = \lim_n P_X((a, b - 1/n]) = F_X(b^-) - F_X(a).$$

### 3.3 The binomial distribution

*Definition 3.5* A random variable over a probability space is called a Bernoulli variable if its range is the set  $\{0, 1\}$ . The distribution of a Bernoulli variable  $X$  is determined by a single parameter  $p_X(1) := p$ . In fact, a Bernoulli variable can be identified with a two-state probability space.

*Definition 3.6* A Bernoulli process is a compound experiment whose constituents are  $n$  independent Bernoulli trials. It is a probability space with sample space

$$\Omega = \{0, 1\}^n,$$

and probability defined on singletons,

$$P(\{(a_1, \dots, a_n)\}) = p^{\text{number of ones}} (1 - p)^{\text{number of zeros}}.$$

Consider a Bernoulli process (this defines a probability space), and set the random variable  $X$  to be the number of “ones” in the sequence (the number of successes out of  $n$  repeated Bernoulli trials). The range of  $X$  is  $\{0, \dots, n\}$ , and its point distribution is

$$p_X(k) = P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}. \quad (3.1)$$

*Definition 3.7* A random variable  $X$  over a probability space  $(\Omega, \mathcal{F}, P)$  is called a binomial variable (משתנה בינומית) with parameters  $(n, p)$  if it takes integer values between zero and  $n$  and its point distribution is (3.1). We write  $X \sim \mathcal{B}(n, p)$ .

*Discussion:* A really important point: one often encounters problems starting with a statement “ $X$  is a binomial random variable, what is the probability that bla bla bla..” without any mention of the underlying probability space  $(\Omega, \mathcal{F}, P)$ . Is this legitimate? There are two answers to this point: (i) if the question only addresses the random variable  $X$ , then it can be fully solved knowing just the distribution  $P_X$ ; the fact that there exists an underlying probability space is irrelevant for the sake of answering this kind of questions. (ii) The triple  $(\Omega = \{0, 1, \dots, n\}, \mathcal{F} = 2^\Omega, P = P_X)$  is a legitimate probability space. In this context the random variable  $X$  is the identity map  $X(x) = x$ .

*Example:* Diapers manufactured by Pamp-ggies are defective with probability 0.01. Each diaper is defective or not independently of other diapers. The company sells diapers in packs of 10. The customer gets his/her money back only if *more than one* diaper in a pack is defective. What is the probability for this to happen?

Every time the customer takes a diaper out of the pack, he faces a Bernoulli trial. The sample space is  $\{0, 1\}$  (1 is defective) with  $p(1) = 0.01$  and  $p(0) = 0.99$ . The number of defective diapers  $X$  in a pack of ten is a binomial variable  $\mathcal{B}(10, 0.01)$ . The probability that  $X$  be larger than one is

$$\begin{aligned} P_X(\{2, 3, \dots, 10\}) &= 1 - P_X(\{0, 1\}) \\ &= 1 - p_X(0) - p_X(1) \\ &= 1 - \binom{10}{0} (0.01)^0 (0.99)^{10} - \binom{10}{1} (0.01)^1 (0.99)^9 \\ &\approx 0.0043. \end{aligned}$$

▲ ▲ ▲

*Example:* An airplane engine breaks down during a flight with probability  $1 - p$ . An airplane lands safely only if *at least half* of its engines are functioning upon landing. What is preferable: a two-engine airplane or a four-engine airplane (or perhaps you'd better walk)?

The number of functioning engines is a binomial variable, in one case  $X_1 \sim \mathcal{B}(2, p)$  and in the second case  $X_2 \sim \mathcal{B}(4, p)$ . The question is whether  $P_{X_1}(\{1, 2\})$  is larger than  $P_{X_2}(\{2, 3, 4\})$  or the other way around. Now,

$$P_{X_1}(\{1, 2\}) = \binom{2}{1}p^1(1-p)^1 + \binom{2}{2}p^2(1-p)^0$$

$$P_{X_2}(\{2, 3, 4\}) = \binom{4}{2}p^2(1-p)^2 + \binom{4}{3}p^3(1-p)^1 + \binom{4}{4}p^4(1-p)^0.$$

Opening the brackets,

$$P_{X_1}(\{1, 2\}) = 2p(1-p) + p^2 = 2p - p^2$$

$$P_{X_2}(\{2, 3, 4\}) = 6p^2(1-p)^2 + 4p^3(1-p) + p^4 = 3p^4 - 8p^3 + 6p^2.$$

One should prefer the four-engine airplane if

$$p(3p^3 - 8p^2 + 7p - 2) > 0,$$

which factors into

$$p(p-1)^2(3p-2) > 0,$$

and this holds only if  $p > 2/3$ . That is, the higher the probability for a defective engine, less engines should be used. ▲ ▲ ▲

Everybody knows that when you toss a fair coin 100 times it will fall Head 50 times... well, at least we know that 50 is the most probable outcome. How probable is in fact this outcome?

*Example:* A fair coin is tossed  $2n$  times, with  $n \gg 1$ . What is the probability that the number of Heads equals exactly  $n$ ?

The number of Heads is a binomial variable  $X \sim \mathcal{B}(2n, \frac{1}{2})$ . The probability that  $X$  equals  $n$  is given by

$$p_X(n) = \binom{2n}{n} \left(\frac{1}{2}\right)^n \left(\frac{1}{2}\right)^n = \frac{(2n)!}{(n!)^2 2^{2n}}.$$

To evaluate this expression we use Stirling's formula,  $n! \sim \sqrt{2\pi n} n^{n+1/2} e^{-n}$ , thus,

$$p_X(n) \sim \frac{\sqrt{2\pi}(2n)^{2n+1/2} e^{-2n}}{2^{2n} 2\pi n^{2n+1} e^{-2n}} = \frac{1}{\sqrt{\pi n}}$$

For example, with a hundred tosses ( $n = 50$ ) the probability that exactly half are Heads is approximately  $1/\sqrt{50\pi} \approx 0.08$ . ▲ ▲ ▲

We conclude this section with a simple fact about the point distribution of a Binomial variable:

*Proposition 3.3* Let  $X \sim \mathcal{B}(n, p)$ , then  $p_X(k)$  increases until it reaches a maximum at  $k = \lfloor (n+1)p \rfloor$ , and then decreases.


*Proof:* Consider the ratio  $p_X(k)/p_X(k-1)$ ,

$$\frac{p_X(k)}{p_X(k-1)} = \frac{n!(k-1)!(n-k+1)! p^k (1-p)^{n-k}}{k!(n-k)! n! p^{k-1} (1-p)^{n-k+1}} = \frac{(n-k+1)p}{k(1-p)}.$$

$p_X(k)$  is increasing if

$$(n-k+1)p > k(1-p) \quad \Rightarrow \quad (n+1)p - k > 0.$$

■

 *Exercise 3.2* In a sequence of Bernoulli trials with probability  $p$  for success, what is the probability that  $a$  successes will occur before  $b$  failures? (Hint: the issue is decided after at most  $a+b-1$  trials).

## 3.4 The Poisson distribution

*Definition 3.8* A random variable  $X$  is said to have a Poisson distribution with parameter  $\lambda$ , if it takes values  $S = \{0, 1, 2, \dots\}$ , and its point distribution is

$$p_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}.$$

(Prove that this defines a probability distribution.) We write  $X \sim \mathcal{Poi}(\lambda)$ .

The first question any honorable person should ask is “why”? After all, we can define infinitely many distributions, and give them fancy names. The answer is that certain distributions are important because they frequently occur in real life. The Poisson distribution appears abundantly in life, for example, when we measure the number of radio-active decays in a unit of time. In fact, the following analysis reveals the origins of this distribution.

*Comment:* Remember the inattentive secretary. When the number of letters  $n$  is large, we saw that the probability that exactly  $k$  letters reach their destination is approximately a Poisson variable with parameter  $\lambda = 1$ .

Consider the following model for radio-active decay. Every  $\epsilon$  seconds (a very short time) a single decay occurs with probability proportional to the length of the time interval:  $\lambda\epsilon$ . With probability  $1 - \lambda\epsilon$  no decay occurs. Physics tells us that this probability is independent of history. The number of decays in one second is therefore a binomial variable  $X \sim \mathcal{B}(n = 1/\epsilon, p = \lambda\epsilon)$ . Note how as  $\epsilon \rightarrow 0$ ,  $n$  goes to infinity and  $p$  goes to zero, but their product remains finite. The probability of observing  $k$  decays in one second is

$$\begin{aligned} p_X(k) &= \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k n(n-1) \dots (n-k+1)}{k! n^k} \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k}{k!} \left(1 - \frac{1}{n}\right) \dots \left(1 - \frac{k-1}{n}\right) \frac{\left(1 - \frac{\lambda}{n}\right)^n}{\left(1 - \frac{\lambda}{n}\right)^k}. \end{aligned}$$

Taking the limit  $n \rightarrow \infty$  we get

$$\lim_{n \rightarrow \infty} p_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}.$$

Thus the Poisson distribution arises from a Binomial distribution when the probability for success in a single trial is very small but the number of trials is very large such that their product is finite.


*Example:* Suppose that the number of typographical errors in a page is a Poisson variable with parameter  $1/2$ . What is the probability that there is at least one error?

This exercise is here mainly for didactic purposes. As always, we need to start by constructing a probability space. The data tells us that the natural space to take is

the sample space  $\Omega = \{0\} \cup \mathbb{N}$  with a probability  $p(k) = e^{-1/2}/(2^k k!)$ . Then the answer is

$$P(\{k \in \mathbb{N} : k \geq 1\}) = 1 - p(0) = 1 - e^{-1/2} \approx 0.395.$$

While this is a very easy exercise, note that we converted the data about a “Poisson variable” into a probability space over the natural numbers with a Poisson distribution. Indeed, a random variable *is* a probability space. ▲ ▲ ▲

 *Exercise 3.3* Assume that the number of eggs laid by an insect is a Poisson variable with parameter  $\lambda$ . Assume, furthermore, that every egg has a probability  $p$  to develop into an insect. What is the probability that exactly  $k$  insects will survive? If we denote the number of survivors by  $X$ , what kind of random variable is  $X$ ? (Hint: construct first a probability space as a compound experiment).

### 3.5 The Geometric distribution

Consider an infinite sequence of Bernoulli trials with parameter  $p$ , i.e.,  $\Omega = \{0, 1\}^{\mathbb{N}}$ , and define the random variable  $X$  to be the number of trials until the first success is met. This random variable takes values in the set  $S = \{1, 2, \dots\}$ . The probability that  $X$  equals  $k$  is the probability of having first  $(k - 1)$  failures followed by a success:

$$p_X(k) = P_X(\{k\}) = P(X = k) = (1 - p)^{k-1} p.$$

A random variable having such a point distribution is said to have a *geometric distribution* with parameter  $p$ ; we write  $X \sim \text{Geo}(p)$ .

*Comment:* The number of failures until the success is met, i.e.,  $X - 1$ , is also called a geometric random variable. We will stick to the above definition.

*Example:* There are  $N$  white balls and  $M$  black balls in an urn. Each time, we take out one ball (with replacement) until we have a black ball. (1) What is the probability that we need  $k$  trials? (2) What is the probability that we need at least  $n$  trials.

The number of trials  $X$  is distributed  $\text{Geo}(M/(M + N))$ . (1) The answer is simply

$$\left(\frac{N}{M + N}\right)^{k-1} \frac{M}{M + N} = \frac{N^{k-1} M}{(M + N)^k}.$$

(2) The answer is

$$\frac{M}{M+N} \sum_{k=n}^{\infty} \left( \frac{N}{M+N} \right)^{k-1} = \frac{M}{M+N} \frac{\left( \frac{N}{M+N} \right)^{n-1}}{1 - \frac{N}{M+N}} = \left( \frac{N}{M+N} \right)^{n-1},$$

which is obviously the probability of failing the first  $n - 1$  times.

▲ ▲ ▲

An important property of the geometric distribution is its lack of memory (משתנה חסר זיכרון). That is, the probability that  $X = n$  given that  $X > k$  is the same as the probability that  $X = n - k$  (if we know that we failed the first  $k$  times, it does not imply that we will succeed earlier when we start the  $k + 1$ -st trial, that is

$$\underbrace{P_X(\{n\} \mid \{k+1, \dots\})}_{P(X=n \mid X>k)} = \underbrace{p_X(n-k)}_{P(X=n-k)}.$$

This makes sense even if  $n \leq k$ , provided we extend  $P_X$  to all  $\mathbb{Z}$ . To prove this claim we follow the definitions. For  $n > k$ ,

$$\begin{aligned} P_X(\{n\} \mid \{k+1, k+2, \dots\}) &= \frac{P_X(\{n\} \cap \{k+1, \dots\})}{P_X(\{k+1, k+2, \dots\})} \\ &= \frac{P_X(\{n\})}{P_X(\{k+1, k+2, \dots\})} \\ &= \frac{(1-p)^{n-1}p}{(1-p)^k} = (1-p)^{n-k-1}p = p_X(n-k). \end{aligned}$$

(14 hrs)  (14 hrs)

### 3.6 The negative-binomial distribution

A coin with probability  $p$  for Heads is tossed until a total of  $n$  Heads is obtained. Let  $X$  be the number of failures until  $n$  successes were met. We say that  $X$  has the *negative-binomial distribution* with parameters  $(n, p)$ . What is  $p_X(k)$  for  $k = 0, 1, 2, \dots$ ? The answer is simply

$$p_X(k) = \binom{n+k-1}{k} p^n (1-p)^k.$$

This is a special instance of negative-binomial distribution, which can be extended to non-integer  $n$ . To allow for non-integer  $n$  we introduce the  $\Gamma$ -function:

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt.$$

Using integration by parts,

$$\Gamma(x+1) = \int_0^{\infty} t^x e^{-t} dt = \int_0^{\infty} x t^{x-1} e^{-t} dt = x \Gamma(x).$$

Since  $\Gamma(1) = 1$ , it follows that

$$\Gamma(2) = 1, \quad \Gamma(3) = 2 \cdot 1,$$

and so on.

Thus, for integers,  $\Gamma(n) = (n-1)!$ . The general negative-binomial distribution with parameters  $0 < p < 1$  and  $r > 0$  has the point distribution,

$$p_X(k) = \frac{\Gamma(r+k)}{k! \Gamma(r)} p^r (1-p)^k.$$

We write  $X \sim n\text{Bin}(r, p)$ .

### 3.7 Other examples

*Example:* Here is a number theoretic result derived by probabilistic means. Recall that the harmonic series is divergent,

$$\sum_{n=1}^{\infty} \frac{1}{n} = \infty.$$

On the other hand, both the geometric series

$$\sum_{n=1}^{\infty} \frac{1}{2^n} = 1,$$

and the series of inverse squares

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$$



converge. Note that both these series include a subset of the terms of the harmonic series, which diverges. In other words, the sequence  $2^n$  and  $n^2$  grow “sufficiently fast”, so that the series of their inverse converge. We now pose the following question: does the series of inverse primes,

$$\sum_{p \text{ prime}} \frac{1}{p} \quad \text{converge?}$$

If the answer is positive, then it would mean that the prime numbers become sparse “sufficiently fast”; in the opposite case, it would mean that their density decays “slowly”. It was Euler who first proved in 1737 that the inverse prime series diverges. Euler’s proof was not really rigorous. Sound proofs were given in forthcoming years, notably by Erdős.

We will now prove that the inverse prime series diverges, using a probabilistic framework. Let  $s > 1$  and let  $X$  be a random variable taking values in  $\mathbb{N}$  with point distribution,

$$p_X(k) = \frac{k^{-s}}{\zeta(s)},$$

where  $\zeta$  is the *Riemann zeta-function*,

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}.$$

Note that  $\zeta(s)$  is well-defined for every  $s > 1$  (e.g., using Cauchy’s condensation test). Moreover,

*Lemma 3.1*

$$\lim_{s \rightarrow 1} \zeta(s) = \infty.$$

*Proof:* Let  $M \in \mathbb{R}$ . Since the harmonic series diverges, there exists an  $N$  such that

$$\sum_{n=1}^N \frac{1}{n} > M + 1.$$

Since by limit arithmetic,

$$\lim_{s \rightarrow 1} \sum_{n=1}^N \frac{1}{n^s} = \sum_{n=1}^N \frac{1}{n} > M + 1,$$

for all  $s$  sufficiently close to 1,

$$\sum_{n=1}^N \frac{1}{n^s} > M.$$

Thus,  $\zeta(s) > M$ , which completes the proof. ■

Let  $A_m \subset \mathbb{N}$  be the set of integers divisible by  $m$ . Clearly,

$$P_X(A_m) = \frac{\sum_{k \text{ divisible by } m} k^{-s}}{\sum_{n=1}^{\infty} n^{-s}} = \frac{\sum_{k=1}^{\infty} (mk)^{-s}}{\sum_{n=1}^{\infty} n^{-s}} = \frac{1}{m^s}.$$

Next, we show that the events  $A_p$ , with  $p$  prime are independent. Indeed,  $A_p \cap A_q = A_{pq}$ , so that

$$P_X(A_p \cap A_q) = P_X(A_{pq}) = \frac{1}{(pq)^s} = \frac{1}{p^s} \cdot \frac{1}{q^s} = P_X(A_p)P_X(A_q).$$

The same consideration holds for all collections of  $A_p$ .

Next, we define for  $q$  prime

$$B_q = \bigcap_{p \text{ prime} \leq q} A_p^c,$$

which is the set of integers that do not have a prime divisor less than  $q$  (for example,  $B_5$  includes all of the integers that are not divisible by 2, 3 and 5).  $(B_q)$  is a decreasing sequence satisfying

$$\bigcap_{q \text{ prime}} B_q = \{1\}.$$

By the continuity of probability for decreasing events,

$$p_X(1) = P_X\left(\bigcap_{q \text{ prime}} B_q\right) = \lim_{q \rightarrow \infty} P_X(B_q) = \lim_{q \rightarrow \infty} \prod_{p \text{ prime} \leq q} P_X(A_p^c) = \prod_{p \text{ prime}} P_X(A_p^c),$$

i.e.,

$$\frac{1}{\zeta(s)} = \prod_{p \text{ prime}} \left(1 - \frac{1}{p^s}\right),$$

an identity known as *Euler's formula*.

Taking the logarithm,

$$-\log \zeta(s) = \sum_{p \text{ prime}} \log \left(1 - \frac{1}{p^s}\right).$$

Since  $p \geq 2$  and  $s > 1$ , it follows that  $0 < 1/p^s < 0.5$ . It can easily be checked that for  $0 < x < 0.5$ ,  $\log(1 - x) \geq -2x$ , hence

$$\log \zeta(s) \leq 2 \sum_{p \text{ prime}} \frac{1}{p^s},$$

namely

$$\frac{1}{2} \log \zeta(s) \leq \sum_{p \text{ prime}} \frac{1}{p^s} \leq \zeta(s).$$

Letting  $s \rightarrow 1$  we conclude that the *harmonic prime series* diverges. ▲ ▲ ▲

(15 hrs)  (15 hrs)

### 3.8 Jointly-distributed random variables

Consider a probability space  $(\Omega, \mathcal{F}, P)$  and a pair of random variables,  $X$  and  $Y$ . That is, we have two maps between probability spaces:

$$\begin{aligned} (\Omega, \mathcal{F}, P) &\xrightarrow{X} (S_X, \mathcal{F}_X, P_X) \\ (\Omega, \mathcal{F}, P) &\xrightarrow{Y} (S_Y, \mathcal{F}_Y, P_Y). \end{aligned}$$

Recall that the probability that  $X$  be in a set  $A \in \mathcal{F}_X$  is fully determined by the distribution  $P_X$ . Now, try to answer the following question: suppose that we are only given the distributions  $P_X$  and  $P_Y$  (i.e., we don't know  $P$ ). What is the probability that  $X(\omega) \in A$  and  $Y(\omega) \in B$ , where  $A \in \mathcal{F}_X$  and  $B \in \mathcal{F}_Y$ ? We cannot answer this question because the knowledge of  $P_X$  amounts to knowing only the probability of events of the form

$$\{\omega \in \Omega : X(\omega) \in A\},$$

with  $A \in \mathcal{F}_X$ , whereas the knowledge of  $P_Y$  amounts to knowing only the probability of events of the form

$$\{\omega \in \Omega : Y(\omega) \in B\},$$

with  $B \in \mathcal{F}_Y$ . Events of the form

$$\{\omega \in \Omega : X(\omega) \in A, Y(\omega) \in B\},$$

are not in the union of these two classes of events. The knowledge of the *separate* distributions of  $X$  and  $Y$  is insufficient to answer questions about events that are *joint* to  $X$  and  $Y$ .

The correct way to think about a pair of random variables is as a mapping  $\Omega \rightarrow S_X \times S_Y$ , i.e.,

$$\omega \mapsto (X(\omega), Y(\omega)).$$

We equip  $S_X \times S_Y$  with a  $\sigma$ -algebra of events  $\mathcal{F}_{X,Y}$  and we require that every set  $A \in \mathcal{F}_{X,Y}$  has a pre-image in  $\mathcal{F}$ . In fact, given the  $\sigma$ -algebra  $\mathcal{F}_{X,Y}$ , the  $\sigma$ -algebra  $\mathcal{F}_X$  is a restriction of  $\mathcal{F}_{X,Y}$ ,

$$\mathcal{F}_X = \{A \subseteq S_X : A \times S_Y \in \mathcal{F}_{X,Y}\},$$

and similarly for  $\mathcal{F}_Y$ .

The *joint distribution* (התפלגות משותפת) of the pair  $X, Y$  is defined naturally as

$$P_{X,Y}(A) := P(\{\omega \in \Omega : (X(\omega), Y(\omega)) \in A\}).$$

One can infer the *marginal distributions* (התפלגויות שוליות) of  $X$  and  $Y$  from this joint distribution, as

$$\begin{aligned} P_X(A) &= P_{X,Y}(A \times S_Y) & A \in \mathcal{F}_X \\ P_Y(B) &= P_{X,Y}(S_X \times B) & B \in \mathcal{F}_Y. \end{aligned}$$

When both  $S_X$  and  $S_Y$  are countable spaces, we define the *joint point distribution*,

$$p_{X,Y}(x, y) := P_{X,Y}(\{(x, y)\}) = P(X = x, Y = y).$$

Obviously,

$$\begin{aligned} p_X(x) &= P_{X,Y}(\{x\} \times S_Y) = \sum_{y \in S_Y} p_{X,Y}(x, y) \\ p_Y(y) &= P_{X,Y}(S_X \times \{y\}) = \sum_{x \in S_X} p_{X,Y}(x, y). \end{aligned}$$

Finally, we define the *joint cumulative distribution function*,

$$F_{X,Y}(x, y) := P_{X,Y}((-\infty, x] \times (-\infty, y]) = P(X \leq x, Y \leq y).$$

*Example:* There are three red balls, four white balls and five blue balls in an urn. We extract three balls. Let  $X$  be the number of red balls and  $Y$  the number of white balls. What is the joint distribution of  $X$  and  $Y$ ?

The natural probability space here is the set of triples out of twelve elements,

$$\Omega = \{(i, j, k) : 1 \leq i < j < k \leq 12\},$$

endowed with uniform probability. Then,

$$(X, Y) : \Omega \rightarrow \{(i, j) : i, j \geq 0, i + j \leq 3\}.$$


For example,

$$(X, Y)((3, 4, 5)) = (1, 2).$$

Then,

$$p_{X,Y}(0, 0) = \frac{\binom{5}{3}}{\binom{12}{3}} \quad \text{and} \quad p_{X,Y}(1, 1) = \frac{\binom{3}{1}\binom{4}{1}\binom{5}{1}}{\binom{12}{3}},$$

etc. ▲ ▲ ▲

 **Exercise 3.4** Construct two probability spaces, and on each define two random variables,  $X, Y$ , such that the two  $P_X$  are the same and the two  $P_Y$  are the same, but the  $P_{X,Y}$  differ.

**Multiple random variables** These notions can be easily generalized to  $n$  random variables.  $X_1, \dots, X_n$  are viewed as a function from  $\Omega$  to the product set  $S_1 \times \dots \times S_n$ , with joint distribution

$$P_{X_1, \dots, X_n}(A) = P(\{\omega : (X_1(\omega), \dots, X_n(\omega)) \in A\}),$$

where  $A \in S_1 \times \dots \times S_n$ . The *marginal distributions* of subsets of variables are obtained, for example,

$$P_{X_1, \dots, X_{n-1}}(A) = P_{X_1, \dots, X_n}(A \times S_n),$$

with  $A \subseteq S_1 \times S_2 \times \dots \times S_{n-1}$ .

### 3.9 Independence of random variables

Let  $(\Omega, \mathcal{F}, P)$  be a probability space and let  $X : \Omega \rightarrow S$  be a random variable, where the set  $S$  is equipped with its  $\sigma$ -algebra of events  $\mathcal{F}_S$ . By the very definition


of a random variable, for every event  $A \in \mathcal{F}_S$ , the event  $X^{-1}(A)$  is an element of  $\mathcal{F}$ . That is,

$$X^{-1}(\mathcal{F}_S) = \{X^{-1}(A) : A \in \mathcal{F}_S\} \subseteq \mathcal{F}.$$

We have seen that  $X^{-1}(\mathcal{F}_S)$  is a  $\sigma$ -algebra, i.e., a sub- $\sigma$ -algebra of  $\mathcal{F}$ . We call it *the  $\sigma$ -algebra generated by the random variable  $X$* , and denote it by  $\sigma(X)$ . Events in  $\sigma(X)$  are subsets of  $\Omega$  (not of  $S$ ) that characterize the outcome of  $X(\omega)$ .

Similarly, when we have a pair of random variables  $X, Y$  with a  $\sigma$ -algebra  $\mathcal{F}_{X,Y}$ , they generate (together!) a  $\sigma$ -algebra,  $\sigma(X, Y)$ , which consists of all events of the form

$$\{\omega \in \Omega : (X(\omega), Y(\omega)) \in A\},$$

 **Exercise 3.5** Let  $X$  be a random variable (a measurable mapping) from  $(\Omega, \mathcal{F}, P)$  to the space  $(S, \mathcal{F}_S, P_X)$ . Consider the collection of events,

$$\{X^{-1}(A) : A \in \mathcal{F}_S\},$$

which is by assumption a subset of  $\mathcal{F}$ . Prove that this collection is a  $\sigma$ -algebra.

We are now ready to define the independence of two random variables. Recall that we already have a definition for the independence of events:

***Definition 3.9** Two random variables  $X, Y$  over a probability space  $(\Omega, \mathcal{F}, P)$  are said to be independent if every event in  $\sigma(X)$  is independent of every event in  $\sigma(Y)$ . In other words, they are independent if every information associated with the value of  $X$  does not affect the (conditional) probability of events reflecting only the random variable  $Y$ .*

***Example:*** Consider the probability space associated with tossing two dice, and let  $X$  be the sum of the dice and  $Y$  be the value of the first die, i.e.,

$$X((i, j)) = i + j \quad \text{and} \quad Y((i, j)) = i.$$

The ranges of  $X$  and  $Y$  are  $S_X = \{2, \dots, 12\}$  and  $S_Y = \{1, \dots, 6\}$ , respectively. The  $\sigma$ -algebras generated by  $X$  and  $Y$  are

$$\begin{aligned} \sigma(X) &= \{\{(i, j) \in \Omega : i + j \in A\} : A \subseteq \{2, \dots, 12\}\} \\ \sigma(Y) &= \{\{(i, j) \in \Omega : i \in B\} : B \subseteq \{1, \dots, 6\}\}. \end{aligned}$$

Recall that the events  $X^{-1}(\{7\})$  and  $Y^{-1}(\{3\})$  are independent. Does it mean that  $X$  and  $Y$  are independent variables? No, for example  $X^{-1}(\{6\})$  and  $Y^{-1}(\{3\})$  are

dependent. It is not true that *any* information on the outcome of  $X$  does not change the probability of the outcome of  $Y$ . ▲ ▲ ▲

While the definition of independence may seem hard to work with, it is easily translated into simpler terms. Let  $A \times B$  be an event in  $\mathcal{F}_{X,Y}$  with  $A \in \mathcal{F}_X$  and  $B \in \mathcal{F}_Y$ . If  $X$  and  $Y$  are independent, then

$$\begin{aligned} P_{X,Y}(A \times B) &= P(X \in A, Y \in B) \\ &= P(X \in A)P(Y \in B) \\ &= P_X(A)P_Y(B). \end{aligned}$$


In particular, if  $A = \{x\}$  and  $B = \{y\}$  are singletons, then

$$p_{X,Y}(x, y) = p_X(x)p_Y(y).$$

Finally, if  $A = (-\infty, x]$  and  $B = (-\infty, y]$ , then

$$F_{X,Y}(x, y) = F_X(x)F_Y(y).$$

Thus, two random variables are independent only if their joint distribution (joint point distribution, joint cumulative distribution function) factors into a product of distributions.

 *Exercise 3.6* Prove that two random variables  $X, Y$  are independent *if and only if*

$$P_{X,Y}(A \times B) = P_X(A)P_Y(B)$$

for every  $A \in \mathcal{F}_X$  and  $B \in \mathcal{F}_Y$ .

These definitions are easily generalized to  $n$  random variables. The random variables  $X_1, \dots, X_n$  have a joint distribution  $P_{X_1, \dots, X_n}$  defined on a  $\sigma$ -algebra of events of the form

$$\{\omega \in \Omega : (X_1(\omega), \dots, X_n(\omega)) \in A\}, \quad A \in S_1 \times \dots \times S_n.$$

These variables are mutually independent if for all  $A_1 \in \mathcal{F}_1, \dots, A_n \in \mathcal{F}_n$ ,

$$P_{X_1, \dots, X_n}(A_1 \times \dots \times A_n) = P_{X_1}(A_1) \dots P_{X_n}(A_n).$$

We further extend the definition to a countable number of random variables. An infinite sequence of random variables is said to be mutually independent if every finite subset is independent.

We will see now a strong use of independence. But first an important lemma, which is the “second half” of a lemma whose first part we have already seen. Recall the *first Borel-Cantelli lemma* that states that if an infinite sequence of events  $(A_n)$  has the property that  $\sum_n P(A_n) < \infty$ , then

$$P(\limsup_n A_n) = 0.$$

There is also a converse lemma, which however requires the independence of the events:

*Lemma 3.2 (Second Borel-Cantelli)* Let  $(A_n)$  be a sequence of mutually independent events in a probability space  $(\Omega, \mathcal{F}, P)$ . If  $\sum_n P(A_n) = \infty$ , then

$$P(\limsup_n A_n) = 1.$$

*Proof:* Note that

$$(\limsup_n A_n)^c = \left( \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k \right)^c = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k^c = \liminf_n A_n^c.$$


Fix  $n$ . Because the events are independent we have

$$P\left(\bigcap_{k=n}^{\infty} A_k^c\right) = \prod_{k=n}^{\infty} (1 - P(A_k)).$$

Using the inequality  $1 - x \leq e^{-x}$ ,

$$P\left(\bigcap_{k=n}^{\infty} A_k^c\right) \leq \prod_{k=n}^{\infty} e^{-P(A_k)} = \exp\left(-\sum_{k=n}^{\infty} P(A_k)\right) = 0,$$

where we have used the divergence of the series. Thus, the event  $(\limsup_n A_n)^c$  is a countable union of events that have zero probability, and therefore also has zero probability. It follows that its complement has probability one. ■

 *Exercise 3.7* Show, by means of a counter example, why does the second Borel-Cantelli lemma require the independence of the random variables.



*Example:* Consider an infinite sequence of Bernoulli trials with probability  $0 < p < 1$  for “success”. What is the probability that the sequence SFS appears infinitely many times? Let  $A_j$  be the event that the sub-sequence  $a_j a_{j+1} a_{j+2}$  equals SFS, i.e.,

$$A_j = \{(a_n) \in \{S, F\}^{\mathbb{N}} : a_j = S, a_{j+1} = F, a_{j+2} = S\}.$$

The events  $A_1, A_4, A_7, \dots$  are independent. Since they have an equal positive probability,  $p^2(1-p)$ ,

$$\sum_{n=1}^{\infty} P(A_{3n}) = \infty \quad \Rightarrow \quad P(\limsup_n A_{3n}) = 1.$$

▲ ▲ ▲

*Example:* Here is a more subtle application of the second Borel-Cantelli lemma. Let  $(X_n)$  be an infinite sequence of independent random variables assuming real positive values, and having the following cumulative distribution function,

$$F_X(x) = \begin{cases} 0 & x \leq 0 \\ 1 - e^{-x} & x > 0 \end{cases}.$$

(Such random variables are called exponential; we shall study them later on). Thus, for any positive  $x$ ,

$$P(X_j > x) = e^{-x}.$$

In particular, we may ask about the probability that the  $n$ -th variable exceeds  $\alpha \log n$ ,

$$P(X_n > \alpha \log n) = e^{-\alpha \log n} = n^{-\alpha}.$$

It follows from the two Borel-Cantelli lemmas that

$$P(X_n > \alpha \log n \text{ i.o.}) = \begin{cases} 0 & \alpha > 1 \\ 1 & \alpha \leq 1 \end{cases}.$$

By the same method, we can obtain refined estimates, such as

$$P(X_n > \log n + \alpha \log \log n \text{ i.o.}) = \begin{cases} 0 & \alpha > 1 \\ 1 & \alpha \leq 1 \end{cases},$$

and so on.

▲ ▲ ▲

### 3.10 Sums of random variables

Let  $X, Y$  be two (discrete) real-valued random variables with joint distribution  $P_{X,Y}$ . Let  $Z = X + Y$ , that is, if

$$X : \Omega \rightarrow S_X \subset \mathbb{R} \quad \text{and} \quad Y : \Omega \rightarrow S_Y \subset \mathbb{R},$$

then

$$Z : \Omega \rightarrow S_X + S_Y = \{x + y : x \in X, y \in Y\} \equiv S_Z$$

is given by

$$Z(\omega) = X(\omega) + Y(\omega).$$

Note that for every  $S_Z$ ,

$$Z^{-1}(\{z\}) = \bigcup_{x \in S_X} X^{-1}(\{x\}) \cap Y^{-1}(\{z - x\}),$$

so that

$$p_Z(z) = P(Z^{-1}(\{z\})) = \sum_{x \in S_X} P(X^{-1}(\{x\}) \cap Y^{-1}(\{z - x\})) = \sum_{x \in S_X} p_{X,Y}(x, z - x).$$

For the particular case where  $X$  and  $Y$  are independent we have

$$p_{X+Y}(z) = \sum_{x \in S_X} p_X(x) p_Y(z - x),$$


the last expression being the *discrete convolution* of  $p_X$  and  $p_Y$  evaluated at the point  $z$ .

*Example:* Let  $X \sim \text{Poi}(\lambda_1)$  and  $Y \sim \text{Poi}(\lambda_2)$  be independent random variables. What is the distribution of  $X + Y$ ?

Using the convolution formula, and the fact that Poisson variables assume non-negative integer values,

$$\begin{aligned} p_{X+Y}(k) &= \sum_{j=0}^k p_X(j) p_Y(k-j) \\ &= \sum_{j=0}^k e^{-\lambda_1} \frac{\lambda_1^j}{j!} e^{-\lambda_2} \frac{\lambda_2^{k-j}}{(k-j)!} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{k!} \sum_{j=0}^k \binom{k}{j} \lambda_1^j \lambda_2^{k-j} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{k!} (\lambda_1 + \lambda_2)^k, \end{aligned}$$

i.e., the sum of two *independent* Poisson variables is a Poisson variable, whose parameter is the sum of the two parameters. ▲ ▲ ▲

 **Exercise 3.8** Let  $X \sim \mathcal{B}(n, p)$  and  $Y \sim \mathcal{B}(m, p)$ . Prove that  $X+Y \sim \mathcal{B}(n+m, p)$ . Give an intuitive explanation for why this must hold.

### 3.11 Conditional distributions

Recall our definition of the conditional probability: if  $A$  and  $B$  are events, then

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

This definition embodies the notion of prediction that  $A$  has occurred given that  $B$  has occurred. We now extend the notion of conditioning to random variables:

*Definition 3.10* Let  $X, Y$  be (discrete) random variables over a probability space  $(\Omega, \mathcal{F}, P)$ . We denote their joint point distribution by  $p_{X,Y}$ ; it is a function  $S_X \times S_Y \rightarrow [0, 1]$ . The conditional point distribution (התפלגות נקודתית מותנה) of  $X$  given  $Y$  is defined as

$$p_{X|Y}(x \mid y) := P(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{p_{X,Y}(x, y)}{p_Y(y)}.$$

(It is defined only for values of  $y$  for which  $p_Y(y) > 0$ .)

*Example:* Let  $p_{X,Y}$  be defined by the following table:

$Y, X$	0	1
0	0.4	0.1
1	0.2	0.3

What is the conditional distribution of  $X$  given  $Y$ ?

Answer:

$$\begin{aligned}
 p_{X|Y}(0 | 0) &= \frac{p_{X,Y}(0, 0)}{p_Y(0)} = \frac{0.4}{0.4 + 0.1} \\
 p_{X|Y}(1 | 0) &= \frac{p_{X,Y}(1, 0)}{p_Y(0)} = \frac{0.1}{0.4 + 0.1} \\
 p_{X|Y}(0 | 1) &= \frac{p_{X,Y}(0, 1)}{p_Y(1)} = \frac{0.2}{0.2 + 0.3} \\
 p_{X|Y}(1 | 1) &= \frac{p_{X,Y}(1, 1)}{p_Y(1)} = \frac{0.3}{0.2 + 0.3}.
 \end{aligned}$$

▲ ▲ ▲


Note that we always have

$$p_{X,Y}(x, y) = p_{X|Y}(x | y)p_Y(y).$$

Summing over all  $y \in S_Y$ ,

$$p_X(x) = \sum_{y \in S_Y} p_{X,Y}(x, y) = \sum_{y \in S_Y} p_{X|Y}(x | y)p_Y(y),$$

which can be identified as the *law of total probability* formulated in terms of random variables.

 *Exercise 3.9* True or false: every two random variables  $X, Y$  satisfy

$$\begin{aligned}
 \sum_{x \in S_X} p_{X|Y}(x | y) &= 1 \\
 \sum_{y \in S_Y} p_{X|Y}(x | y) &= 1.
 \end{aligned}$$

*Example:* Assume that the number of eggs laid by an insect is a Poisson variable with parameter  $\lambda$ . Assume, furthermore, that every eggs has a probability  $p$  to develop into an insect. What is the probability that exactly  $k$  insects will survive?

This problem has been previously given as an exercise. We will solve it now in terms of conditional distributions. Let  $X$  be the number of eggs laid by the insect, and  $Y$  the number of survivors. We don't even bother to (explicitly) write the probability space, because we have all the needed data as distributions and conditional distributions. We know that  $X$  has a Poisson distribution with parameter  $\lambda$ , i.e.,

$$p_X(n) = e^{-\lambda} \frac{\lambda^n}{n!} \quad n = 0, 1, \dots,$$

whereas the distribution of  $Y$  conditional on  $X$  is binomial,

$$p_{Y|X}(k | n) = \binom{n}{k} p^k (1-p)^{n-k} \quad k = 0, 1, \dots, n.$$

The distribution of the number of survivors  $Y$  is then

$$\begin{aligned} p_Y(k) &= \sum_{n=0}^{\infty} p_{Y|X}(k | n) p_X(n) = \sum_{n=k}^{\infty} \binom{n}{k} p^k (1-p)^{n-k} e^{-\lambda} \frac{\lambda^n}{n!} \\ &= e^{-\lambda} \frac{(\lambda p)^k}{k!} \sum_{n=k}^{\infty} \frac{[\lambda(1-p)]^{n-k}}{(n-k)!} \\ &= e^{-\lambda} \frac{(\lambda p)^k}{k!} e^{\lambda(1-p)} = e^{-\lambda p} \frac{(\lambda p)^k}{k!}. \end{aligned}$$

Thus,  $Y \sim \text{Poi}(\lambda p)$ . ▲ ▲ ▲

(17 hrs)  (17 hrs)

*Example:* Let  $X \sim \text{Poi}(\lambda_1)$  and  $Y \sim \text{Poi}(\lambda_2)$  be *independent* random variables. What is the conditional distribution of  $X$  given that  $X + Y = n$ ?

We start by writing things explicitly,

$$\begin{aligned} p_{X|X+Y}(k | n) &= P(X = k | X + Y = n) \\ &= \frac{P(X = k, X + Y = n)}{P(X + Y = n)} \\ &= \frac{p_{X,Y}(k, n-k)}{\sum_{j=0}^n p_{X,Y}(j, n-j)}. \end{aligned}$$

At this point we use the fact that the variables are independent and their distributions are known:

$$\begin{aligned} p_{X|X+Y}(k | n) &= \frac{e^{-\lambda_1} \frac{\lambda_1^k}{k!} e^{-\lambda_2} \frac{\lambda_2^{n-k}}{(n-k)!}}{\sum_{j=0}^n e^{-\lambda_1} \frac{\lambda_1^j}{j!} e^{-\lambda_2} \frac{\lambda_2^{n-j}}{(n-j)!}} \\ &= \frac{\binom{n}{k} \lambda_1^k \lambda_2^{n-k}}{\sum_{j=0}^n \binom{n}{j} \lambda_1^j \lambda_2^{n-j}} \\ &= \frac{\binom{n}{k} \lambda_1^k \lambda_2^{n-k}}{(\lambda_1 + \lambda_2)^n}. \end{aligned}$$

Thus, it is a binomial distribution with parameters  $n$  and  $\lambda_1/(\lambda_1 + \lambda_2)$ , which we may write as

$$[X \text{ conditional on } X + Y = n] \sim \mathcal{B}\left(n, \frac{\lambda_1}{\lambda_1 + \lambda_2}\right).$$

▲ ▲ ▲

**Generalization** Conditional probabilities can be generalized to multiple variables. For example,

$$p_{X,Y|Z}(x, y | z) := P(X = x, Y = y | Z = z) = \frac{p_{X,Y,Z}(x, y, z)}{p_Z(z)}$$

$$p_{X|Y,Z}(x | y, z) := P(X = x | Y = y, Z = z) = \frac{p_{X,Y,Z}(x, y, z)}{p_{Y,Z}(y, z)},$$

and so on.

*Proposition 3.4* Every three random variables  $X, Y, Z$  satisfy

$$p_{X,Y,Z}(x, y, z) = p_{X|Y,Z}(x | y, z) p_{Y|Z}(y | z) p_Z(z).$$

*Proof:* Immediate. Just follow the definitions. ■

*Example:* Consider a sequence of random variables  $(X_k)_{k=0}^n$ , each assuming values in a finite alphabet  $\mathcal{A} = \{1, \dots, s\}$ . Their joint distribution can be expressed as follows:

$$p_{X_1, \dots, X_n}(x_0, x_1, \dots, x_n) = p_{X_n|X_1, \dots, X_{n-1}}(x_n | x_0, \dots, x_{n-1})$$

$$p_{X_{n-1}|X_0, \dots, X_{n-2}}(x_{n-1} | x_0, \dots, x_{n-2}) \dots p_{X_1|X_0}(x_1 | x_0) p_{X_0}(x_0).$$

There exists a class of such sequences called *Markov chains* (שרשראות מרקוב). In a Markov chain,

$$p_{X_n|X_1, \dots, X_{n-1}}(x_n | x_0, \dots, x_{n-1}) = p_{X_n|X_{n-1}}(x_n | x_{n-1}),$$

i.e., the distribution of  $X_n$  “depends on its history only through its predecessor”; if  $X_{n-1}$  is known, then the knowledge of its predecessors is superfluous for the sake of predicting  $X_n$ . Note that this does not mean that  $X_n$  is independent of  $X_{n-2}$ ! Furthermore, a Markov chain is said to be *time-homogeneous* if the functions  $p_{X_k | X_{k-1}}(x | y)$  are the same for all  $k$ , i.e., can be represented by an  $s$ -by- $s$  matrix,  $M_{xy}$ .


Thus, for a Markov chain,

$$\begin{aligned} p_{X_1, \dots, X_n}(x_0, x_1, \dots, x_n) &= p_{X_n | X_{n-1}}(x_n | x_{n-1}) p_{X_{n-1} | X_{n-2}}(x_{n-1} | x_{n-2}) \dots p_{X_1 | X_0}(x_1 | x_0) p_{X_0}(x_0) \\ &= M_{x_n, x_{n-1}} M_{x_{n-1}, x_{n-2}} \dots M_{x_1, x_0} p_{X_0}(x_0). \end{aligned}$$

If we now sum over all values that  $X_0$  through  $X_{n-1}$  can assume, then

$$p_{X_n}(x_n) = \sum_{x_{n-1} \in \mathcal{A}} \dots \sum_{x_0 \in \mathcal{A}} M_{x_n, x_{n-1}} M_{x_{n-1}, x_{n-2}} \dots M_{x_1, x_0} p(x_0) = \sum_{x_0 \in \mathcal{A}} M_{x_n, x_0}^n p_{X_0}(x_0).$$

Thus, the distribution on  $X_n$  is related to the distribution of  $X_0$  through the application of the  $n$ -power of a matrix, the *transition matrix* (מטריצת המעבר). Situations of interest are when the distribution of  $X_n$  tends to a limit, which does not depend on the initial distribution of  $X_0$ . Such Markov chains are said to be *ergodic*. When the rate of approach to this limit is exponential, the Markov chain is said to be *exponentially mixing*. ▲ ▲ ▲

 **Exercise 3.10** Let  $X, Y, Z$  be three random variables. We say that  $X$  is independent of  $Z$  given  $Y$  if

$$p_{X|YZ}(x | y, z) = p_{X|Y}(x | z)$$

for all  $z$ . Is this relation symmetric? Does it imply that  $Y$  is independent of  $X$  given  $Z$ ? Does it imply that  $X$  is independent of  $Y$  given  $Z$ .