CHAPTER 12

DISCRETE PROBABILITIES

Probability theory is a fundamental tool in many domains and, in particular computer science, where its main use is of course the average case analysis of algorithms. Combinatorics allow us to count the number of elements of a set, to count the number of operations of an algorithm on given data. However, when the set or the data are subject to change, the number of elements or operations can also change, and probabilities come into the picture to study average values, the deviation with respect to these average values, etc. Probabilities are also used in other areas in computer science, e.g.

- probabilistic algorithms: a random choice can be on average, or even almost always, more efficient than computing the exact choice,
- modelling and simulation,
- queueing theory: for instance, study the average waiting time for accessing the nerve-centres of a network,
- signal processing,
- probabilistic arguments sometimes allow us to prove properties of algorithms which are not provable otherwise.

This chapter defines the notions of discrete probability distribution, conditional probability distribution and independance, and random variables. The chapter reviews Bayes's rule and some applications, the weak law of large numbers, how to use generating series to study random variables and the main properties of the most common probability distributions.

We strongly recommend the following classic handbook:

William Feller, *Probability Theory*, Vol. 1, John Wiley, New York (1968).

12.1 Generalities

12.1.1 Terminology

Probability theory studies randomness; e.g.:

- the outcome of a coin-tossing game,
- the number of daily calls through a telephone switchboard,
- the waiting time for accessing a network,
- the length of the life span (without breakdowns) of an operating system.

Probability theory describes a mathematical model for such random experiments, where 'random' means 'depending on chance'.

12.1.2 Sample spaces

The first basic notion is the notion of *trial, or sample point or observation*: a trial is the outcome of a random experiment; such an outcome is denoted by ω and the set of all possible trials is traditionally denoted by Ω and called the *sample space* i.e.: each outcome obtained in a possible experiment will correspond to a unique element ω in Ω. The fulfillment of the random experiment is thus reduced to the random choice of ω in Ω. Of course, different experiments will lead to different sets Ω.

Remark 12.1 Note that the same 'experiments' in the intuitive sense may correspond to different random experiments according to what we are interested in. For instance, assuming a coin-tossing game with a dime and a quarter, and assuming that the coins never fall on their edges:

- If we are interested in the outcome Heads or Tails, the sample space will be the set of possible outcomes, i.e. $\Omega = \{HH, HT, TH, TT\}$.
- If we are interested in the number of Heads, the sample space will be $\Omega = \{0, 1, 2\}$.
- If we are interested in the concordance of the outcomes, i.e. in the value of the predicate 'both coins fell on the same side', we will have $\Omega = \{$Agreement, Disagreement$\}$.

Example 12.2

1. Consider the random experiment consisting of rolling a pair of dice (we will often use this example); the possible outcomes are pairs of integers $\omega = (x, y)$, with $1 \leq x, y \leq 6$; the sample space is thus the set $\Omega = \{1, 2, \ldots, 6\}^2$.
2. In the case of the random experiment consisting of counting the number of daily calls through a telephone switchboard, the possible outcomes are: $\{1, 2, \ldots, n, \ldots\}$. The sample space will thus be $\Omega = \mathbb{N}$.

12.1.3 Events

The second basic notion is the notion of *event*, i.e. an event whose fulfillment depends upon the outcome of a random experiment. We will thus represent an event A by the set of outcomes of the experiments fulfilling it, and we will identify

$$A = \{\omega \mathbin{/} A \text{ occurs if } \omega \text{ is the outcome of the experiment}\}\,.$$

Example 12.3
1. When throwing two dice, event A: 'the total score is ≥ 10' is represented in $\Omega = \{1, \ldots, 6\}^2$ by the set of pairs (x, y) such that $x + y \geq 10$, i.e.

$$A = \{(4,6), (5,6), (6,6), (5,5), (6,5), (6,4)\}\,.$$

2. When counting the number of phone calls, event A: 'the switchboard can put through at most 5000 daily calls' is represented by $A = \{n \in \mathbb{N}/n \leq 5000\}$.

12.1.4 Relations among events

We briefly describe the logical operations which can be performed on events; events being represented by subsets of the sample space Ω, these logical operations will correspond to the usual set-theoretical operations, up to the terminology.

Each event A is associated with its *complementary event*, denoted by A^c (or $\neg A$, or $\overline{A}$): event A^c occurs if and only if A does not occur, and is represented in Ω by the complement of the subset representing A.

Example 12.4 In throwing two dice, the event 'the total score is less than 10' is represented in $\Omega = \{1, \ldots, 6\}^2$ by the 30 $(= 36 - 6)$ pairs (x, y) such that $x + y < 10$, which is the complement of the set A representing the event 'the total score is ≥ 10'.

For each pair of events A_1 and A_2, event 'A_1 *and* A_2' (resp. 'A_1 *or* A_2') occurs if and only if both A_1 and A_2 occur (resp. either A_1 or A_2 or both occur). In the space Ω, A_1 *and* A_2 (resp. A_1 *or* A_2) is represented by the intersection (resp. the union) of A_1 and A_2. Subsequently, we will denote A_1 *and* A_2 (resp. A_1 *or* A_2) by $A_1 \cap A_2$ (resp. $A_1 \cup A_2$). *The impossible event*, denoted by $\emptyset$, is represented by the empty set in Ω. Two events A_1 and A_2 are said to be *mutually exclusive* if and only if $A_1 \cap A_2 = \emptyset$; if A_1 and A_2 are mutually exclusive, then A_1 *and* A_2 is impossible, and A_1 *or* A_2 is denoted by $A_1 + A_2$ (disjoint union). Events $\{A_n \mathbin{/} n \in \mathbb{N}\}$ are said to be *mutually exclusive* if they are pairwise mutually exclusive. The *sure event*, which always occurs whatever the outcome of the experiment, is represented by Ω. Finally, event A is said to *imply* event B if, whenever A occurs, B also occurs, i.e. the subset of Ω representing A is contained in the subset of Ω representing B.

Example 12.5 Consider again the tossing of two dice. Define the two events A_1 'the total is ≥ 10', and A_2 'the total is < 10'. These two events are mutually exclusive and event 'A_1 *or* A_2' is sure. (In general, if two events A_1 and A_2 are mutually exclusive, the event 'A_1 *or* A_2' is not sure.)

A set $\{A_n \ / \ n \in \mathbb{N}\}$ of mutually exclusive events such that $\Omega = \cup_{n\in\mathbb{N}} A_n$ is said to be a *partition* of Ω . Summing up, the correspondence between probabilistic and set-theoretic terminologies is shown in the following table:

Probabilistic terminology	Set-theoretic terminology	Notation
Sure event	Whole space	Ω
Impossible event	Empty set	$\emptyset$
Complementary event	Complementary subset	A^c ($\overline{A}$, $\neg A$)
And	Intersection	$\cap$
Or	Union	$\cup$
Mutually exclusive events	Disjoint subsets	$A_1 \cap A_2 = \emptyset$
Partition	Partition	$\sum A_i = \Omega$, with $A_i \cap A_j = \emptyset$
Implication	Inclusion	$A \subseteq B$

12.2 Probability spaces

12.2.1 Probability space

We have seen that a random experiment can be described by a set Ω (the sample space), and a class $\mathcal{T}$ of subsets of Ω (the subsets representing events). When Ω is not denumerable, $\mathcal{T}$ cannot be the set $\mathcal{P}(\Omega)$ of all subsets of Ω: $\mathcal{P}(\Omega)$ will indeed contain too many pathological subsets which cannot represent natural events. In order to properly represent the notion of event a class $\mathcal{T}$ of subsets should, 1. contain the sure event Ω, 2. be stable by complement, 3. be stable by denumerable unions and intersections, which gives:

Definition 12.6 *A class $\mathcal{T}$ of subsets of Ω is said to be a tribe if it satisfies*

(i) $\Omega \in \mathcal{T}$,
(ii) $A \in \mathcal{T} \implies A^c \in \mathcal{T}$,
(iii) $\forall n \in \mathbb{N}$, $A_n \in \mathcal{T} \implies \bigcup_{n\in\mathbb{N}} A_n \in \mathcal{T}$.

Condition (iii) *is called the σ-additivity principle, or the additivity principle if A_n, $n = 1, \dots, p$ is finite. A tribe is also called a σ-algebra, or a σ-additive class or a countably additive class.*

Remark 12.7 $\mathcal{T} = \mathcal{P}(\Omega)$ always satisfies conditions (i)–(iii); often, for a finite or denumerable Ω, we will have $\mathcal{T} = \mathcal{P}(\Omega)$.

Remark 12.8 Since $\mathcal{T}$ is stable by complementation and denumerable union, it also is stable by denumerable intersection, indeed

$$\bigcap_{n \in \mathbb{N}} A_n = \left(\bigcup_{n \in \mathbb{N}} A_n^c \right)^c .$$

A pair $(\Omega, \mathcal{T})$ formed by a set and a tribe on this set gives a model describing random experiments and the corresponding events. However, the goal of probability theory is not simply to find a descriptive model, but mainly to supply a tool for computing quantitative evaluations. That is what we will now do.

The notion of *probability* corresponds intuitively to the notion of 'likelihood or frequency of realization' of an event A during a sequence of repetitions of a random experiment in which A may (or may not) occur. If N is the number of repetitions of a random experiment and N_A the number of realizations of A during a sequence of N experiments, the ratio N_A/N tends to become stable when N grows. The probability of A is the limit of the ratio N_A/N when $N \to \infty$. (This notion will appear again in the law of large numbers.) Some properties are satisfied by the mapping $A \mapsto N_A/N$:

(i) $0 \leq N_A/N \leq 1$,

(ii) if $A_1, \dots, A_N$ are mutually exclusive,

$$N_{A_1 \cup \dots \cup A_n}/N = N_{A_1}/N + \dots + N_{A_n}/N .$$

Hence the following definition.

Definition 12.9

1. *Let $\mathcal{T}$ be a tribe in Ω. A probability distribution P on $(\Omega, \mathcal{T})$ is a mapping $P\colon \mathcal{T} \longrightarrow [0,1]$ such that*

(i) $\forall A \in \mathcal{T}, \quad 0 \leq P(A) \leq 1$,

(ii) $P(\Omega) = 1$,

(iii) *if $(A_n)_{n \in \mathbb{N}}$ is a family such that: for all i, j in $\mathbb{N}$, $A_i, A_j \in \mathcal{T}$ and $A_i \cap A_j = \emptyset$, then*

$$P\left(\bigcup_{n \in \mathbb{N}} A_n \right) = \sum_{n \in \mathbb{N}} P(A_n) .$$

2. *When Ω is denumerable and $\mathcal{T} = \mathcal{P}(\Omega)$, one can substitute for the above conditions $\forall\omega \in \Omega$, $0 \leq P(\omega) \leq 1$, $P(A) = \sum_{\omega\in A} P(\omega)$ and $\sum_{\omega\in\Omega} P(\omega) = 1$, where $P(\omega)$ is an abbreviation for $P(\{\omega\})$.*

Exercise 12.1 Verify that in case 2 of the preceding definition, the mapping P: $\mathcal{P}(\Omega) \longrightarrow [0, 1]$ defined by $P(A) = \sum_{\omega\in A} P(\omega)$ is a probability distribution. ♢

If Ω is a finite space and $\mathcal{T} = \mathcal{P}(\Omega)$, the *uniform distribution* on Ω is often considered: it gives each elementary event ω the same probability, $p = \dfrac{1}{|\Omega|}$, since $\sum_{\omega\in\Omega} P(\omega) = \sum_{\omega\in\Omega} p = |\Omega|p = 1$ must hold. Each sample point ω is said to be equally probable, and the uniform distribution corresponds to the intuitive notion of 'random choice'. We then have, for all events A, $P(A) = \dfrac{|A|}{|\Omega|}$ (the ratio of the number of favourable cases to the total number of cases).

Example 12.10 Consider again the tossing of two dice, and endow the sample space Ω with the uniform distribution. Each possible outcome $\omega = (x, y)$ has probability $1/36$, and event A ='the total is ≥ 10' has probability $P(A) = 6/36 = 1/6$.

If Ω is an infinite space there can no longer be a uniform distribution.
In all cases, Ω finite or infinite, an event A is said to be

- *almost sure* if it occurs *almost surely*, i.e. $P(A) = 1$,
- *almost impossible* if it almost surely will not occur, i.e. $P(A) = 0$.

Example 12.11 Consider the random experiment consisting of tossing a coin till Tails turns up, then stopping. The space Ω is described by the sequence of outcomes, where Tails (resp. Heads) is abbreviated in T (resp. H).

$$\Omega = \{T, HT, HHT, \ldots, H^nT, \ldots, HHH\cdots\} = \Omega' \cup \{HHH\cdots\}\ .$$

Assume that, throwing the coin once, we have $P(T) = p$ and $P(H) = q = 1 - p$, with $0 < p < 1$. Then $P(H^nT) = q^np$ (this will be formally proved later on), and

$$\sum_{\omega\in\Omega'} P(\omega) = \sum_{n\in\mathbb{N}} q^np = p\sum_{n\in\mathbb{N}} q^n = \frac{p}{1-q} = 1.$$

The event $A = HHH\cdots$, i.e. an infinite sequence of Heads as outcome, is a logically possible event, corresponding to a non-empty subset of $\mathcal{P}(\Omega)$, but it is almost impossible since $P(A) = 0$.

Let us conclude with some useful formulas

Proposition 12.12
(i) *For $A_1, \ldots, A_n$ mutually exclusive events,*

$$P(A_1 \cup \cdots \cup A_n) = \sum_{i=1}^{n} P(A_i)\,.$$

(ii) $\forall A, B, \quad P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
(iii) $\forall A, \quad P(A^c) = 1 - P(A)$.
(iv) $\forall A, B, \qquad A \subseteq B \quad \Longrightarrow \quad P(A) \leq P(B)$.
(v) $\forall A_1, \ldots, A_n\,, \quad P(\cup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} P(A_i)$.
(vi) *Let $(A_n)_{n\in\mathbb{N}}$ be a monotone increasing sequence, i.e. $\forall n$, $A_n \subseteq A_{n+1}$, then* $P(\cup_{n\in\mathbb{N}} A_n) = \lim_{n\to\infty} P(A_n)$.
(vii) *Let $(A_n)_{n\in\mathbb{N}}$ be a monotone decreasing sequence, i.e. $\forall n$, $A_{n+1} \subseteq A_n$, then* $P(\cap_{n\in\mathbb{N}} A_n) = \lim_{n\to\infty} P(A_n)$.

The proof is straightforward.

Proposition 12.13 *Let $A_1, \ldots, A_n$ be events, we have*

$$P(\cup_{i=1}^{n} A_i) = \sum_{k=1}^{n} (-1)^{k-1} \sum_{i_1 < \cdots < i_k} P(A_{i_1} \cap \cdots \cap A_{i_k})\,.$$

Proof. This result is proved in the same way as Proposition 6.13 (counting techniques). □

Exercise 12.2 Two (not so good) snipers A and B take aim at a target. A hits the target with probability 1/4 and B hits the target with probability 2/5. What is the probability that the target is hit? ♢

12.3 Conditional probabilities and independent events

12.3.1 Intuition

Let us start with an introductory example. Consider a space Ω consisting of a population of $N = |\Omega|$ people, with uniform distribution; the corresponding experiment is the 'random' choice of an individual. Among the N people, let H be the subset of men and D be the subset of colourblind people. Let D_e (resp. H_e) be the event: 'the chosen person is colourblind (resp. male)'; we have

$$P(D_e) = \frac{|D|}{|\Omega|} \text{ and } P(H_e) = \frac{|H|}{|\Omega|}\,.$$

Assuming now that we are only interested in the subpopulation of males, and that we randomly choose one man, the probability that he is colourblind is $\dfrac{|H \cap D|}{|H|}$, i.e. the probability of being colourblind for men. It is denoted by $P(D_e/H_e)$, and we say: *'probability of D_e assuming (or given, or on the hypothesis) H_e'*. In the present example, we have $P(D_e/H_e) = \dfrac{P(D_e \cap H_e)}{P(H_e)}$ because $\dfrac{|H \cap D|}{|H|} = \Big(\dfrac{|H \cap D|}{|\Omega|}\Big)\Big/\Big(\dfrac{|H|}{|\Omega|}\Big)$.

We will see that this situation can be generalized to any subpopulation defined by an arbitrary event A. We can always consider a probability distribution restricted to the set of sample points satisfying event A; we then say the *'conditional probability given A'*.

12.3.2 Definitions

Definition 12.14 *Let $(\Omega, \mathcal{T}, P)$ be a probability space and A an event having a positive probability; the conditional probability of event B given A is defined by*

$$P(B/A) = \frac{P(A \cap B)}{P(A)}\,. \tag{12.1}$$

Proposition 12.15 *The mapping $B \mapsto P(B/A)$ is a probability distribution on the space $(\Omega, \mathcal{T})$, for each choice of the conditioning event A.*

Proof. We check that

(i) $0 \le P(B/A) \le 1$, since $P(A \cap B) \le P(A)$,

(ii) $P(\Omega/A) = \dfrac{P(\Omega \cap A)}{P(A)} = \dfrac{P(A)}{P(A)} = 1$,

(iii) for each family $(B_i)_{i \in \mathbb{N}}$ of mutually disjoint subsets, the family $(B_i \cap A)_{i \in \mathbb{N}}$ is also formed of mutually disjoint subsets, thus:

$$\begin{aligned} P\big((\cup_{i \in \mathbb{N}} B_i)/A\big) &= \frac{P\big(\cup_{i \in \mathbb{N}} (B_i \cap A)\big)}{P(A)} = \sum_{i \in \mathbb{N}} \frac{P(B_i \cap A)}{P(A)} \\ &= \sum_{i \in \mathbb{N}} P(B_i/A)\,. \end{aligned}$$ □

Remark 12.16

(1) Equation (12.1) is often used in the form $P(A \cap B) = P(B/A)P(A)$.

(2) Let $(\Omega, \mathcal{P}(\Omega))$ be a finite space together with the uniform distribution, and $A \subseteq \Omega$ be such that $|A| \neq 0$, then $P(B/A) = \dfrac{|B \cap A|}{|A|}$, i.e. the uniform distribution restricted to A.

Example 12.17 We again return to Example 12.11: toss a coin till the first Tails occurs. Let B be the event : 'the first Tails occurs after at least three tosses', and A: 'Tails does not occur at the first toss'; note that $B \subseteq A$. Recall that $p = P(T)$ and $q = P(H)$, where Tails (resp. Heads) is abbreviated to T (resp. H). We have

$$P(B) = \sum_{n\geq 2} q^n p = q^2 p \frac{1}{1-q} \quad \text{and} \quad P(A) = \sum_{n\geq 1} q^n p = qp\frac{1}{1-q}\,,$$

hence

$$P(B/A) = \frac{P(A\cap B)}{P(A)} = \frac{P(B)}{P(A)} = q\,.$$

Example 12.18 Consider families with two children. Letting F (resp. M) stand for girl (resp. boy), and denoting the children by decreasing age, the sample space is $\Omega = \{MM, MF, FM, FF\}$; Ω is endowed with the uniform distribution. Let A be the event 'there is at least one boy' and B the event 'there are two boys'. Assuming there is at least one boy, we compute the probability that there are two boys: $A = \{MM, MF, FM\}$ and $B = \{MM\} = A \cap B$, hence $P(B/A) = 1/3$.

Equation (12.1) can be generalized as follows

Proposition 12.19 (generalization) *Let $A_1, \ldots, A_n$ be arbitrary events in $(\Omega, \mathcal{T})$, then*

$$P(A_1\cap\cdots\cap A_n) = P(A_1)P(A_2/A_1)P(A_3/A_1\cap A_2)\cdots P(A_n/A_1\cap\cdots\cap A_{n-1})\,,$$

provided that we define the right-hand side to be 0 as soon as one of its factors is 0.

Proof. By induction on n;

- Basis: for $n = 2$ we come across the definition of conditional probabilities when $P(A_1) \neq 0$, and the result is clear when $P(A_1) = 0$.
- Inductive step: assuming the result true for n, let us prove it for $n + 1$:
  (i) $P(A_1 \cap\cdots\cap A_{n+1}) > 0$, then we also have $P(A_1\cap\cdots\cap A_n) > 0$, and thus by the definition of conditional probabilities

$$P(A_1\cap\cdots\cap A_{n+1}) = P(A_1\cap\cdots\cap A_n)P(A_{n+1}/A_1\cap\cdots\cap A_n)$$

  by the induction hypothesis we can replace $P(A_1\cap\cdots\cap A_n)$ by $P(A_1)$ $P(A_2/A_1)P(A_3/A_1\cap A_2)\cdots P(A_n/A_1\cap\cdots\cap A_{n-1})$, hence the result.

(ii) $P(A_1 \cap \cdots \cap A_{n+1}) = 0$. Note that the sequence

$$p_i = P(A_1 \cap \cdots \cap A_i)$$

is monotone decreasing; hence

(ii.1) either $P(A_1) \neq 0$, and $\exists i \in \{2, \ldots, n+1\}$, $P(A_1 \cap \cdots \cap A_{i-1}) > 0$ and $P(A_1 \cap \cdots \cap A_i) = 0$; but then $P(A_i/A_1 \cap \cdots \cap A_{i-1}) = 0$ by the definition of conditional probabilities, and the right-hand side is thus 0.

(ii.2) or $P(A_1) = 0$, and both sides are 0. □

Exercise 12.3 An urn contains b black balls and r red balls; k random drawings are performed, $k < r$, with the following rule: if a black ball is drawn it is replaced, if a red ball is drawn c black balls are added. What is the probability of drawing k red balls? ♢

Exercise 12.4 Message passing. Consider passing a message 'yes' or 'no' in a population. Each person passes the message he/she received with probability p and the opposite message with probability $q = 1 - p$. Let X_n be the message received by the nth individual I_n. Assume that I_0 passes 'yes' to I_1. What is the probability that I_n receives 'yes'? ♢

Example 12.20 (Samplings without replacement) An urn contains $N > 2$ balls, b black balls and r red balls, $b + r = N$. Two balls are drawn and not replaced. What is the probability of drawing two red balls? Let C be the event: 'two red balls were drawn', A_1 be the event: 'the first drawn ball is red' and A_2: 'the second ball is red'. We clearly have $P(A_1) = \dfrac{r}{N}$, $P(A_2/A_1) = \dfrac{r-1}{N-1}$ and

$$P(C) = P(A_1 \cap A_2) = P(A_1)P(A_2/A_1) = \frac{r \times (r-1)}{N \times (N-1)}\,.$$

This should be likened to the notion of arrangement without repetitions of Chapter 6.

12.3.3 Bayes's rule

It is frequently easier to evaluate conditional probabilities for an event B than to compute its probability $P(B)$. This happens if e.g. the experiment upon which event B depends can be split into partial successive experiments, and if the probability of each partial experiment depends on the outcomes of preceding experiments; see Exercise 12.4 and Exercise 12.6 (probability of causes). There are tools enabling us to deduce $P(B)$ and $P(A_n/B)$ from the $P(B/A_n)$s, and that is what we will now show.

Lemma 12.21 (Total probabilities) *Let $(\Omega, \mathcal{T}, \mathcal{P})$ be a probability space, and let $(A_k)_{k\in\mathbb{N}}$ be a partition; for any event B, we have*

$$P(B) = \sum_{k\in\mathbb{N}} P(B/A_k)P(A_k)\,.$$

Proof. Since $(A_k)_{k\in\mathbb{N}}$ forms a partition, we have $\cup_{k\in\mathbb{N}} A_k = \Omega$; then

$$\begin{aligned} P(B) &= P(B\cap\Omega) = P\big(B\cap(\cup_{k\in\mathbb{N}}A_k)\big) \\ &= \sum_{k\in\mathbb{N}} P(B\cap A_k) \quad \text{(since the } B\cap A_k\text{s are pairwise disjoint)} \\ &= \sum_{k\in\mathbb{N}} P(B/A_k)P(A_k). \end{aligned} \qquad \square$$

This rule can be extended to the case when some $P(A_k)$s are 0, by attributing an arbitrary value to $P(B/A_k)$ and letting $P(B/A_k)P(A_k) = 0$.

Exercise 12.5 (Polya's urn model) An urn contains b black balls and r red balls. A ball is randomly drawn. It is replaced, and, moreover, c balls of the same colour are added. A new random drawing is performed and the whole procedure is iterated. Let $X^n \in \{B, R\} = \Omega$ be the colour of the ball drawn at the nth drawing. Prove by induction on n that $P(X^n = B) = b/(b+r)$ for all n. ♢

The next theorem is known as Bayes's rule for the probability of causes.

Theorem 12.22 *Letting $(\Omega, \mathcal{T}, P)$ be a probability space, $(A_n)_{n\in\mathbb{N}}$ a partition and B an event with positive probability, we have*

$$P(A_n/B) = \frac{P(A_n)P(B/A_n)}{\sum_{k\in\mathbb{N}} P(A_k)P(B/A_k)}\ .$$

Proof. By the definition of conditional probabilities

$$P(A_n/B) = \frac{P(A_n\cap B)}{P(B)} = \frac{P(A_n)P(B/A_n)}{P(B)}$$

and, by the preceding lemma, $P(B) = \sum_{k\in\mathbb{N}} P(A_k)P(B/A_k)$, hence

$$P(A_n/B) = \frac{P(A_n)P(B/A_n)}{\sum_{k\in\mathbb{N}} P(A_k)P(B/A_k)}\ . \qquad \square$$

Exercise 12.6 (Probability of causes) Let U_i, for $i = 1, 2$, be two urns containing r_i red balls and b_i black balls. One urn is randomly chosen, and a ball is drawn from it. Assuming a red ball was drawn, what is the probability that it comes from urn 1? ♢

EXERCISE 12.7 Let A and B be two machines producing, respectively, 100 and 200 objects. A (resp. B) produces 5% (resp. 6%) of flawed objects. Given a flawed object, what is the probability that it was manufactured by A? ♢

EXERCISE 12.8 A city is divided into three areas. During a poll, the repartition of votes for candidate C was as shown in the following table.

Area number	Weight of the area in voters (%)	Score of C in the area (%)
1	30	40
2	50	48
3	20	60

1. What is the probability that a randomly chosen voter voted for C?
2. Given that e voted for C, what is the probability that e is from area 3? ♢

12.3.4 Independent events

Definition 12.23 *Two events A and B in $(\Omega, \mathcal{T}, P)$ are independent if and only if they satisfy the following equivalent conditions:*

(i) $P(B/A) = P(B)$,
(ii) $P(A/B) = P(A)$,
(iii) $P(A \cap B) = P(A)P(B)$.

That these three conditions are equivalent follows immediately from the equality $P(A \cap B) = P(B/A)P(A) = P(A/B)P(B)$.

Proposition 12.24

1. *If A and B are independent, the pairs of events (A, B^c), (A^c, B) and (A^c, B^c) are also independent.*
2. *If A and B are independent, we have:*

$$P(A \cup B) = P(A) + P(B) - P(A)P(B).$$

Proof. Check, for instance, that (A^c, B) and (A^c, B^c) are independent. We have

$$\begin{aligned} P(A^c \cap B) + P(A \cap B) &= P(B) \\ P(A^c \cap B) &= P(B) - P(A \cap B) = P(B) - P(A)P(B) \\ &= P(B)\big(1 - P(A)\big) = P(B)P(A^c); \\ P(A^c \cap B^c) &= 1 - P(A \cup B) = 1 - P(A) - P(B) + P(A)P(B) \\ &= P(A^c) - P(B)\big(1 - P(A)\big) = P(A^c)\big(1 - P(B)\big) \\ &= P(A^c)P(B^c). \end{aligned}$$ □

Example 12.25 Return to Example 12.2 and the tossing of two dice.

(i) Events A = ‘the first die is a one’, and B = ‘the second die is even’, are independent since $P(A) = 1/6$, $P(B) = 1/2$ and $P(A \cap B) = 1/12 = P(A)P(B)$. Note that A and B are not disjoint, since $A \cap B = \{(1,2),(1,4),(1,6)\} \neq \emptyset$.
(ii) Events A = ‘the first die is a one’, and B = ‘the first die is even’ are disjoint, but not independent since $P(A) = 1/6$, $P(B) = 1/2$ and $P(A \cap B) = 0 \neq P(A)P(B)$. In general, disjoint events are never independent, unless one of them has probability 0.

Remember from this example that *the notions of disjoint events and of independent events are orthogonal.*

Exercise 12.9 Consider events A = ‘there is at most one girl in a family’, and B = ‘there are both boys and girls’; check that events A and B are independent if the underlying space Ω_3 is the set of families with three children, but A and B are not independent in the space Ω_2 of families with two children; Ω_2 and Ω_3 are endowed with the corresponding uniform probabilities. ♢

Exercise 12.10 A poll among the students on a course gave the following outcomes:

(a) 2/3 of the students say they like maths, and among them:
- 70% would rather take an exam without notes,
- 20% are in favour of prohibiting smoking.

(b) 1/3 of the students say they do not like maths, and among them:
- 20% would rather take an exam without notes,
- 90% are in favour of prohibiting smoking.

Let M, D, F be the events:
- M= liking maths,
- D= in favour of taking an exam without notes,
- F= in favour of prohibiting smoking,

and $\overline{M}, \overline{D}, \overline{F}$ the complementary events.

Given that a student likes maths, events D and F are independent. Similarly, given that a student does not like maths, events D and F are independent.
1. Compute $P(M)$, $P(D/M)$, $P(F/M)$, $P(\overline{M})$, $P(\overline{D}/M)$, $P(\overline{F}/M)$, $P(\overline{D}/\overline{M})$, $P(D/\overline{M})$, $P(\overline{F}/\overline{M})$, $P(F/\overline{M})$.
2. Given that student e would rather take an exam without notes and is in favour of prohibiting smoking, what is the probability that this student likes maths? ♢

One can generalize the notion of independent events to sets of events.

Definition 12.26 *Let $A_1, \ldots, A_n$ be events in $(\Omega, \mathcal{T}, P)$. $A_1, \ldots, A_n$ are mutually independent if and only if they satisfy the following equivalent conditions*

(i) *For all combinations $1 \leq i_1 < \cdots < i_k \leq n$ we have*

$$P(A_{i_1} \cap \cdots \cap A_{i_k}) = P(A_{i_1}) \cdots P(A_{i_k}) \ .$$

(ii) *For* $1 \leq k \leq n-1$, *and for all combinations* $(i, i_1, \ldots, i_k)$ *of* $k+1$ *pairwise distinct integers among* $\{1, \ldots, n\}$ *we have*

$$P(A_i / A_{i_1} \cap \cdots \cap A_{i_k}) = P(A_i) \; .$$

REMARK 12.27 It is not enough to check that the A_i are pairwise independent, nor that $P(A_1 \cap \cdots \cap A_n) = P(A_1) \cdots P(A_n)$. Consider, for instance, the experiment consisting of throwing a die and looking at the outcome; $\Omega = \{1, 2, \ldots, 6\}$ with the uniform probability. Let A be 'the outcome is ≤ 3', B be 'the outcome is even' and C be 'the outcome is not divisible by 3'. $P(A) = P(B) = 1/2$, $P(C) = 2/3$. $A \cap B \cap C = \{2\}$, hence $P(A \cap B \cap C) = 1/6 = P(A)P(B)P(C)$; but A, B, C are not independent since $P(A \cap B) = 1/6 \neq P(A)P(B)$.

EXERCISE 12.11 Find an example of three events A, B, C, pairwise independent, such that $P(A \cap B \cap C) \neq P(A)P(B)P(C)$. $\diamondsuit$

EXERCISE 12.12 Let the set Ω consist of the eight vertices of a cube constructed on the three unit vectors of origin 0 in a Cartesian space. Each vertex is identified by its three coordinates $i, j, k \in \{0, 1\}$. The probability distribution is defined by

$$P(\omega) = \begin{cases} 0 & \text{if } i+j+k \text{ is even,} \\ \dfrac{1}{4} & \text{otherwise.} \end{cases}$$

1. Prove that this indeed defines a probability distribution.
2. Show that the three events: $A = \{i = 0\}$, $B = \{j = 0\}$, $C = \{k = 0\}$ are pairwise independent.
3. Are events $A \cap B$ and C independent? $\diamondsuit$

12.3.5 Product spaces

Product probability distributions correspond to the notion of repeated random experiments or of a combination of several random experiments. Consider a sequence $(\omega_1, \ldots, \omega_n)$ of random experiments, $\omega_i \in \Omega_i$, for $i = 1, \ldots, n$, and study the global experiment $\omega = (\omega_1, \ldots, \omega_n)$.

Z ω ranges over $\Omega = \Omega_1 \times \cdots \times \Omega_n$. It is natural to endow Ω with a tribe $\mathcal{T}$ of subsets such that all events A consisting of events $A_1, \ldots, A_n$ are in $\mathcal{T}$; however, this seemingly simplest choice for $\mathcal{T}$ which is the set-theoretical product of the $\mathcal{T}_i$ is not suitable, because it is *not* a tribe. We have to choose the least possible tribe $\mathcal{T}$ which is *generated* by this product and which we will denote by $\mathcal{T} = \prod_{i=1}^{n} \mathcal{T}_i$. Lastly, we will define a probability distribution P on $(\Omega, \mathcal{T})$.

Definition 12.28 *Let $(\Omega_i, \mathcal{T}_i, P_i)$ be denumerable probability spaces with $\mathcal{T}_i = \mathcal{P}(\Omega_i)$ for $i = 1, \ldots, n$. The product space $(\Omega, \mathcal{T}, P)$ is defined by*

- $\Omega = \prod_{i=n}^{n} \Omega_i$,
- $\mathcal{T} = \prod_{i=1}^{n} \mathcal{T}_i$, *and thus $\mathcal{T}$ is the least tribe containing $A_1 \times \cdots \times A_n$ for all $A_i \subseteq \Omega_i$, $i = 1, \ldots, n$,*
- $P(\omega_1, \ldots, \omega_n) = P_1(\omega_1) \times \cdots \times P_n(\omega_n)$, $\quad \forall(\omega_1, \ldots, \omega_n) \in \Omega$.

Exercise 12.13 Prove that the above-defined P is indeed a probability distribution. ♢

The same definition can be given for the product of an infinite number of sample spaces.

This definition implicitly assumes that the sequence of represented trials $(\omega_1, \ldots, \omega_n)$ are *independent*, and that is by far the most frequent case in life (repetitions of experiments, sampling populations, etc. . .). Without the independence hypothesis, the only condition that can be demanded of P is that $\forall A_i \in \mathcal{T}_i$,

$$P(\Omega_1 \times \cdots \times \Omega_{i-1} \times A_i \times \Omega_{i+1} \times \cdots \times \Omega_n) = P_i(A_i) \tag{12.2}$$

and this condition is not enough to determine P, as shown by the next example.

Example 12.29 Let $(\Omega_i, \mathcal{T}_i, P_i)$, $i = 1, 2$ be defined by $\Omega_1 = \Omega_2 = \{0, 1\}$, $\mathcal{T}_1 = \mathcal{T}_2 = \mathcal{P}(\Omega_1)$, and let $P_1 = P_2$ be defined by $P_1(0) = P_2(0) = P_1(1) = P_2(1) = 1/2$. We can define two probability distributions P and Q on $(\Omega_1 \times \Omega_2, \mathcal{T}_1 \times \mathcal{T}_2)$

$$P((0,1)) = P((1,0)) = P((0,0)) = P((1,1)) = 1/4\,,$$

$$Q((1,0)) = Q((0,1)) = 1/2\,, \quad Q((0,0)) = Q((1,1)) = 0\,,$$

P and Q both satisfy (12.2); P is a product distribution corresponding to two independent trials.

Example 12.30 (Repeated independent trials. Bernoulli trials) Let $(\Omega, \mathcal{T}, P)$ be the sample space associated with an experiment. Repeat this experiment n times, independently, i.e. the outcome of the ith experiment does not depend upon experiment number j, $j \neq i$. Then, the probability distribution associated with the n successive experiments will be described by the product $(\Omega^n, \mathcal{T}^n, P^n)$. For instance, consider n successive tosses of a coin. We represent Tails by 0 and Heads by 1, and identify {Tails, Heads} by $\mathbb{B}$; with these conventions, the sample space corresponding to the coin-tossing game is $\mathbb{B} = \{0, 1\}$, $\mathcal{T} = \mathcal{P}(\mathbb{B})$, and $P(0) = P(\text{Tails}) = p$, $P(1) = P(\text{Heads}) = q$ with $p + q = 1$. The space corresponding to n successive tosses is $(\mathbb{B}^n, \mathcal{P}(\mathbb{B}^n), P)$ where P is defined for

$\omega \in \mathbb{B}^n$ by $P(\omega) = p^k q^{n-k}$ if $|\omega|_0 = k$ and $|\omega|_1 = n - k$, and where $|\omega|_i$ denotes the number of occurrences of i in ω . A sequence of n trials as given above is called a sequence of *Bernoulli trials.*

Exercise 12.14 Construct the sample space corresponding to n repeated independent trials, where each trial has possible outcomes r_i, $i = 1, \dots, k$, with $P(r_i) = p_i$ and $p_1 + \cdots + p_k = 1$. ♢

Example 12.31 (Samplings with replacement) An urn contains r red balls and b black balls; a ball is drawn and replaced. The experiment is repeated n times: then the probability of drawing a red ball k times and a black ball $(n - k)$ times is

$$\left(\frac{r}{r+b}\right)^k \left(\frac{b}{r+b}\right)^{n-k}.$$

This can be likened to the notion of permutations with repetitions of Chapter 6.

The product spaces are useful in modelling genetics (Mendel's laws), in reliability studies (life span before breakdowns, etc.) for systems.

12.4 Random variables

12.4.1 Definitions

First note that the same experiment, when repeated, does not give the same result each time. This variability is described with random variables. A *random variable* is a function whose value depends upon the outcome of a random experiment, i.e. a function defined on a random space.

Example 12.32 Return to the example of the tossing of two dice: the sum S of the scores of the tossing of the dice is a random variable. Indeed, if $\omega = (m, n)$ is the pair of scores of the toss of the dice, we can write $S(\omega) = m + n$ for $\omega = (m, n)$, $1 \le m, n \le 6$.

Definition 12.33 *Let $(\Omega, \mathcal{T}, P)$ be a sample space. A discrete random variable (abbreviated to r.v.) on $(\Omega, \mathcal{T}, P)$ is a mapping $X\colon \Omega \longrightarrow D$, D denumerable, $D \subseteq \mathbb{R}$, such that $\forall d \in D, \quad X^{-1}(d) \in \mathcal{T}$.*

The condition we demand is that the inverse image of an element of D be in $\mathcal{T}$: this condition is required in order to be able to speak of the probability that X assumes the value d. This probability will be denoted by $P(X^{-1}(d))$, or $P(X = d)$.

Example 12.34 A mapping which is not a random variable. Let the experiment consist of throwing a die; let $\Omega = \{1,2,3,4,5,6\}$, let the tribe $\mathcal{T} = \{\Omega, \emptyset, \text{Even}, \text{Odd}\}$, where Even $= \{2,4,6\}$, Odd $= \{1,3,5\}$ and consider the sample space $(\Omega, \mathcal{T}, P)$, where $P(\Omega) = 1$, $P(\emptyset) = 0$, $P(\text{Even}) = P(\text{Odd}) = 1/2$. Define $X : \Omega \longrightarrow \mathbb{R}$ by

$$X(1) = X(3) = 4 \;, \quad X(2) = X(4) = 6 \;, \quad X(5) = X(6) = 11 \;,$$

then X is not a random variable because $X^{-1}(4) = \{1,3\}$ is not in the tribe $\mathcal{T}$.

Proposition 12.35 *Let $X : \Omega \longrightarrow D$ be an r.v. on $(\Omega, \mathcal{T}, P)$; X defines a probability distribution P_X on $(D, \mathcal{P}(D))$ by*

$$P_X(d) = P(X = d) = P(X^{-1}(d))$$

for $d \in D$. P_X is called the probability distribution of X.

We also define the distribution function of X

$$F_X(d) \;=\; P(X < d) \;=\; P(X^{-1}] - \infty, d[) \;=\; \sum_{d' < d} P_X(d') \,.$$

Remark 12.36 X defines the probability distribution P_X, but P_X, conversely, does not determine X, as shown by the following example. Let X and Y be two r.v.'s $\mathbb{B} \longrightarrow \mathbb{B} \subseteq \mathbb{R}$, having the same distribution

$$P(X = 0) = P(X = 1) = P(Y = 0) = P(Y = 1) = 1/2 \,.$$

Different pairs X, Y will satisfy this requirement, for instance:

- $P_1(X = x, Y = y) = 1/4 \,, \quad \forall x, y \in \mathbb{B} \,.$
- $P_2(X = x, Y = y) = 1/2 \,,$ for $x \neq y$, and $P_2(X = x, Y = x) = 0$ for $x \in \mathbb{B}$.

In the first case (P_1), X and Y are independent, in the second case (P_2), X and Y are dependent, but in both cases X and Y have the same distribution. Giving the distribution of X is thus not enough to fully describe all characteristics of X.

Exercise 12.15 We state some properties of distribution functions that can be verified as an exercise:

1. F is monotone.
2. $\lim_{x \to -\infty} F(x) = 0, \quad \lim_{x \to \infty} F(x) = 1.$ ♢

Definition 12.37 *Let X and Y be two r.v.'s $X : \Omega \longrightarrow D$ and $Y : \Omega \longrightarrow D'$. The function $P_{(X,Y)}(d, d') = P(X = d, Y = d')$ is a probability distribution on $(D \times D', \mathcal{P}(D \times D'))$, called the joint distribution of the r.v.'s X and Y.*

REMARK 12.38 $(X = d, Y = d')$ is an abbreviated notation for the intersection $X^{-1}(d) \cap Y^{-1}(d')$ which is an event of the tribe $\mathcal{T}$ of Ω.

Given the joint distribution of (X, Y) we can determine the distributions of X and Y (called *marginal distributions*); by noting that event $X = d$ can be split into a disjoint union: $(X = d) = \sum_{d' \in D'} (X = d, Y = d')$, thus $P_X(d) = \sum_{d' \in D'} P_{(X,Y)}(d, d')$. Conversely, the marginal distributions of X and Y do not determine the joint distribution of (X, Y), except when the r.v.'s X and Y are independent (see below). The joint distributions $P_{(X,Y)}$ and $Q_{(X,Y)}$ described by the tables given below (see Example 12.29) are different but they correspond to the same marginal distribution $P_X = P_Y = Q_X = Q_Y$.

$P_{(X,Y)}$	x	0	1
	y		
	0	1/4	1/4
	1	1/4	1/4

$Q_{(X,Y)}$	x	0	1
	y		
	0	0	1/2
	1	1/2	0

EXERCISE 12.16 Check that for the r.v.'s X, Y the joint distribution $P_{(X,Y)}$ and $Q_{(X,Y)}$ of the above example (see Example 12.29) are indeed probability distributions, and similarly for the marginal distributions P_X and Q_X. ♢

EXAMPLE 12.39 Let X, Y be two r.v.'s assuming values in $\mathbb{B} = \{0, 1\}$. Assume that the joint distribution of (X, Y) is given by

$$P_{(X,Y)}(0,0) = \frac{1}{8}\,, \qquad P_{(X,Y)}(1,0) = \frac{3}{8}\,,$$
$$P_{(X,Y)}(0,1) = \frac{2}{8}\,, \qquad P_{(X,Y)}(1,1) = \frac{2}{8}\,.$$

The marginal distributions P_X and P_Y of X and Y are given by

$$\begin{aligned} P_X(0) &= \sum_{y \in \mathbb{B}} P_{(X,Y)}(0, y) \\ &= P(X = 0, Y = 0) + P(X = 0, Y = 1) = \frac{3}{8}\,, \end{aligned}$$

and similarly,

$$P_X(1) = \frac{5}{8}\,,\; P_Y(0) = P_Y(1) = \frac{1}{2}\,.$$

Definition 12.40 (Independence of r.v.'s) *Let X and Y be two r.v.'s defined on the same sample space and assuming values in D and D'. X and Y are said to be independent if and only if they verify the following equivalent conditions:*

(i) *$\forall A \subseteq D$, $\forall A' \subseteq D'$, the events $(X \in A)$ and $(Y \in A')$ are independent, where $X \in A$ is an abbreviation for $X^{-1}(A)$.*
(ii) *The joint distribution of (X, Y) is the product of the marginal distributions of X and Y.*
(iii) $\forall d \in D$, $\forall d' \in D'$, $\quad P(X = d, Y = d') = P(X = d)P(Y = d')$.
(iv) $\forall d \in D$, $\forall d' \in D'$, $\quad P(X = d/Y = d') = P(X = d)$.

The implications and equivalences (i) $\Longrightarrow$ (ii) $\Longrightarrow$ (iii) $\Longleftrightarrow$ (iv) are straightforward; to show (iv) $\Longrightarrow$ (i) it is enough to compute $P(X \in A, Y \in A')$.

Example 12.41
(i) The r.v.'s X and Y of Example 12.39 are not independent, because, e.g. $P_X(0) = 3/8, P_Y(0) = 4/8$, but $P_{(X,Y)}(0,0) = 1/8 \neq (3/8) \times (4/8)$.
(ii) Consider the tossing of two dice, and let X be the score of the first die, and Y the score of the second die. X and Y are two r.v.'s defined on $\big(\Omega = (\{1, \ldots, 6\})^2, \mathcal{P}(\Omega)\big)$ together with the uniform distribution into $\{1, \ldots, 6\}$. X and Y are easily seen to be independent, since $\forall i, j,\ 1 \leq i, j \leq 6$, we have

$$P(X = i, Y = j) = 1/36 = (1/6) \times (1/6) = P(X = i) \times P(Y = j)\,.$$

Exercise 12.17 Let X and Y be two independent r.v.'s, $X{:}\,\Omega \longrightarrow D \subseteq \mathrm{I\!R}$ and $Y{:}\,\Omega \longrightarrow D' \subseteq \mathrm{I\!R}$, and f and g two arbitrary functions from $\mathrm{I\!R}$ into $\mathrm{I\!R}$; we can define by composition two new r.v.'s $f(X) = f \circ X$ and $g(Y) = g \circ Y$; are the r.v.'s $f(X)$ and $g(Y)$ independent? ♢

Exercise 12.18 Let W be an r.v. assuming values 1,2,3 with the same probability. Let X, Y and Z be three independent r.v.'s, each with the same distribution as W. Let $U = X + Y$, $V = X - Z$.

1. What are the values assumed by U and by V? What are the distributions of U and V?
2. Write the table giving the probability distribution of the pair (U, V). Are the r.v.'s U and V independent? ♢

Exercise 12.19 How would you define the independence of n r.v.'s ? ♢

Proposition 12.42 *Let $X : \Omega \longrightarrow D$ and $Y : \Omega \longrightarrow D'$ be two independent r.v.'s defined on the same sample space and assuming values in D and D'. Let $f : D \longrightarrow E$ and $g : D' \longrightarrow E'$ be two functions; then $f(X) : \Omega \longrightarrow E$ and $g(Y) : \Omega \longrightarrow E'$ are two independent r.v.'s.*

Proof.

$$\begin{aligned} P(f(X) = e, g(Y) = e') &= P(X \in f^{-1}(e), Y \in g^{-1}(e')) \\ &= P(X \in f^{-1}(e))P(Y \in g^{-1}(e')) \\ &= P(f(X) = e)P(g(Y) = e') \; . \end{aligned}$$ □

12.4.2 Mean and variance of a random variable

We wish to simplify the representation of an r.v. We can say, loosely, that the mean of an r.v. represents the average value of this r.v., and that its variance, or its standard deviation, gives a measure of the error in approximating the r.v. by its mean. For instance, if in a population consisting of n families, we have exactly n_k families with k children, then $P_k = n_k/n$ represents the probability that a 'randomly' chosen family have exactly k children, and the average number of children per family will be $1/n(\sum_{k\geq 0} kn_k)$; if we define the r.v. X as being the number of children of a 'randomly' chosen family, then $1/n(\sum_{k\geq 0} kn_k) = \sum_{k\geq 0} k(n_k/n) = \sum_{k\geq 0} kP_k$, represents the average value of X, i.e. the mean of X. The formal definition follows.

Definition 12.43 *Let X be an r.v.; the mean (also called expectation, average or expected value) of X is defined by*

$$E(X) = \sum_{\omega\in\Omega} X(\omega)P(\omega) = \sum_{d\in D} dP(X = d) \; , \tag{12.3}$$

provided that this sum is defined.

Example 12.44 Return to Example 12.39. We have $E(X) = 5/8$, and similarly $E(Y) = 1/2$. Lastly, we can define the mean of (X, Y) by $E((X, Y)) = (E(X), E(Y))$ and we obtain $E((X, Y)) = (5/8, 1/2)$.

Exercise 12.20 Let Ω be a sample space and $A \subseteq \Omega$; recall that the characteristic function χ_A of A is a mapping $\chi_A\colon \Omega \longrightarrow \mathbb{B} \subseteq \mathbb{R}$ defined by

$$\chi_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{otherwise.} \end{cases}$$

1. Under what condition is χ_A an r.v.?
2. Check that, in this case, $E(\chi_A) = P(A)$. ♢

EXAMPLE 12.45 A computation of the average complexity of an algorithm $\mathcal{A}$ is the computation of a mean. Let $d_1, \ldots, d_k$ be all the data of size n, $P(d_i)$ the probability of datum d_i and $X(d_i)$ the complexity of $\mathcal{A}$ on the datum d_i. The average complexity of $\mathcal{A}$ on data of size n is then given by the mean of the r.v. X i.e. $E(X) = \sum_{i=1}^{k} P(d_i)X(d_i)$.

Proposition 12.46 *Let X and Y be two r.v.'s: $X\colon \Omega \longrightarrow A \subseteq \mathbb{R}$ and $Y\colon \Omega \longrightarrow B \subseteq \mathbb{R}$; we have*

(i) $X \le Y \implies E(X) \le E(Y)$,
(ii) $E(aX + bY) = aE(X) + bE(Y)$,
(iii) *E is linear.*

$X + Y$ (resp aX) is the r.v. defined by $(X + Y)(\omega) = X(\omega) + Y(\omega)$ (resp. $(aX)(\omega) = aX(\omega)$).

Proof. It is straighforward to show that $E(aX) = aE(X)$; similarly (i) is immediate. To check that $E(X+Y) = E(X) + E(Y)$, a computation of multiple sums is necessary; assume $X\colon \Omega \longrightarrow A$, $Y\colon \Omega \longrightarrow B$, and $X + Y\colon \Omega \longrightarrow D$.

$$\begin{aligned}
E(X+Y) &= \sum_{i\in D} iP(X+Y=i) = \sum_{i\in D}\sum_{x+y=i} iP(X=x, Y=y) \\
&= \sum_{x\in A, y\in B} (x+y)P(X=x, Y=y) \\
&= \sum_{(x,y)\in A\times B} xP(X=x, Y=y) + \sum_{(x,y)\in A\times B} yP(X=x, Y=y) \\
&= \sum_{x\in A} x\left(\sum_{y\in B} P(X=x, Y=y)\right) + \sum_{y\in B} y\left(\sum_{x\in A} P(X=x, Y=y)\right) \\
&= \sum_{x\in A} xP(X=x) + \sum_{y\in B} yP(Y=y). \qquad \square
\end{aligned}$$

REMARK 12.47 The mean of an r.v. is not always defined: consider $X\colon \Omega \longrightarrow D$, where $\Omega = \mathbb{N}$, $\mathcal{T} = \mathcal{P}(\mathbb{N})$ and

$$P(X = x) = \begin{cases} 1/2^{n+1} & \text{if } x = 2^n, \\ 0 & \text{otherwise.} \end{cases}$$

X is indeed an r.v. since $\sum_{n\in\mathbb{N}} P(X = x) = 1$, but $E(X)$ is not defined since $\sum_{n\in\mathbb{N}} 2^n/2^{n+1} \longrightarrow \infty$. In technical terms $E(X)$ exists if and only if X is integrable with respect to the measure P_X.

Proposition 12.48 *Let X be an r.v. $\Omega \longrightarrow D$ and f: $D \longrightarrow \mathbb{R}$, then $Y = f(X)$ is an r.v., and $E(Y) = \sum_{d \in D} f(d)P(X = d)$, provided that this sum is defined.*

Proof. Let $D' = f(D)$, then:

$$\begin{aligned} E(Y) &= \sum_{y \in D'} yP(Y = y) = \sum_{y \in D'} yP(f(X) = y) \\ &= \sum_{y \in D'} y(\sum_{\{x \in D / f(x) = y\}} P(X = x)) \\ &= \sum_{y \in D'} (\sum_{\{x \in D / f(x) = y\}} f(x)P(X = x)) \\ &= \sum_{x \in D} f(x)P(X = x)\,. \end{aligned}$$ □

Definition 12.49 *$\forall n \geq 1$, we can define the nth moment of the r.v. X by*

$$m_n(X) = \sum_{x \in D} x^n P(X = x) = E(X^n),$$

provided that this sum is defined.

$E(X)$ is the first moment of X. We now define the *variance* and standard deviation of the r.v. X which, intuitively, measure the distance between X and its mean $E(X)$; i.e. they estimate the fluctuations of X around its mean.

Definition 12.50 *Let X be an r.v. such that $E(X)$ and $E(X^2)$ exist; the variance of X is defined by*

$$var(X) = E((X - E(X))^2) = E(X^2) - (E(X))^2.$$

We define the standard deviation of X by $\sigma(X) = \sqrt{var(X)}$.

It is straightforward to check that

$$\begin{aligned} E((X - E(X))^2) &= E(X^2 - 2XE(X) + E(X)^2) \\ &= E(X^2) - 2E(X)^2 + E(X)^2 = E(X^2) - E(X)^2, \end{aligned}$$

by noting that $E(X)$ is a constant; thus $E(E(X)) = E(X)$.

Proposition 12.51 *Let X be an r.v. and let a and b be constants. Then $var(aX + b) = a^2 var(X)$.*

Proof. We have $var(X + b) = var(X)$, hence the result. □

Definition 12.52 *Let X and Y be two r.v.'s; we define:*

(i) *the co-variance of X and Y by*

$$\Gamma(X,Y) = E[(X-E(X))(Y-E(Y))] = E(XY) - E(X)E(Y) ,$$

(ii) *the correlation coefficient of X and Y by*

$$\rho(X,Y) = \frac{\Gamma(X,Y)}{\sigma(X)\sigma(Y)} .$$

Proposition 12.53 *Let X and Y be two independent r.v.'s, then*

(i) $E(XY) = E(X)E(Y)$,
(ii) $var(X+Y) = var(X) + var(Y)$,
(iii) $\Gamma(X,Y) = \rho(X,Y) = 0$.

Remark 12.54 All converses are *false*, i.e. none of these conditions imply the independence of the r.v.'s X and Y.

Proof. Check, for instance, that (i); (ii) and (iii) are straightforward consequences of (i). Let $X\colon \Omega \longrightarrow A$, $Y\colon \Omega \longrightarrow B$, $XY\colon \Omega \longrightarrow D$:

$$\begin{aligned} E(XY) &= \sum_{xy\in D} xyP(X=x, Y=y) \\ &= \sum_{xy\in D} xyP(X=x)P(Y=y) \quad \text{(since } X \text{ and } Y \text{ are independent)} \\ &= \sum_{x\in A} xP(X=x)(\sum_{y\in B} yP(Y=y)) = \sum_{x\in A} xP(X=x)E(Y) \\ &= E(X)E(Y) . \end{aligned}$$ □

Exercise 12.21
1. What are the means of the r.v.'s U and V defined in Exercise 12.18?
2. What is the correlation coefficient of the r.v.'s U and V defined in Exercise 12.18?

◊

Example 12.55 Consider the two r.v.'s X, Y on $(\mathbb{N}, \mathcal{P}(\mathbb{N})))$ assuming values in $\mathbb{R}$, and such that

$$P(1,0) = P(-1,0) = P(0,1) = P(0,-1) = \frac{1}{4}$$

defines the joint distribution of (X, Y). The marginal distributions are defined by

$$P_X(1) = P_X(-1) = 1/4 \quad P_X(0) = 1/2\ ,$$
$$P_Y(1) = P_Y(-1) = 1/4 \quad P_Y(0) = 1/2\ .$$

Thus $P_X = P_Y = P'$. This can be represented by the following tables:

P x y	-1	0	1
-1	0	1/4	0
0	1/4	0	1/4
1	0	1/4	0

P' x	-1	0	1
	1/4	1/2	1/4

We thus have $E(X) = E(Y) = 0$. Similarly, $E(XY) = 0$, thus also $\rho(X, Y)= 0$. The r.v.'s X and Y thus satisfy all the conditions (i), (ii), (iii) of the preceding proposition, but they are not independent, since, e.g.:

$$P(X = 1, Y = 0) = 1/4 \text{ and } P(X = 1)P(Y = 0) = 1/8.$$

Exercise 12.22

1. Let Z be an r.v. having a geometric distribution (see Section 12.6.3) of ratio a $(0 < a < 1)$, i.e.

$$\forall k \in \mathbb{N}^*, \quad P(Z = k) = a^{k-1}(1 - a)\ .$$

   (a) What is the mean of Z?
   (b) Show that $\forall k \in \mathbb{N}^*$, $P(Z \geq k) = a^{k-1}$.

2. Let X and Y be two independent r.v.'s defined on $(\Omega, \mathcal{T}, P)$, and such that
   X has a geometric distribution of ratio p, $(0 < p < 1)$,
   Y has a geometric distribution of ratio q, $(0 < q < 1)$.

   Define an r.v. T on $(\Omega, \mathcal{T}, P)$, by

$$\forall \omega \in \Omega, \quad T(\omega) = \inf(X(\omega), Y(\omega))\ .$$

   (a) Show that $\forall k \in \mathbb{N}^*$,

$$P(T = k) = P(X \geq k)P(Y \geq k) - P(X \geq k + 1)P(Y \geq k + 1)\ .$$

   (Hint: $P(T = k) = P(T \geq k) - P(T \geq k + 1)$.)

(b) Show that T has a geometric distribution of ratio pq.

3. Consider a sequence $(X_n)_{n\geq 1}$ of independent r.v.'s having the same geometric distribution of ratio p, defined on $(\Omega, \mathcal{T}, P)$. Recall that $(X_n)_{n\geq 1}$ is a sequence of independent r.v.'s if and only if $\forall n \geq 1, \quad X_1, \dots, X_n$ are independent.

For all $n \geq 1$, define an r.v. T_n on $(\Omega, \mathcal{T}, P)$ by

$$\forall \omega \in \Omega\,,\ T_n(\omega) = \inf(X_1(\omega), \dots, X_n(\omega))\,.$$

(a) Show, by induction on n, that T_n has a geometric distribution of ratio p^n.

(b) Show that $\lim_{n\to\infty} P(T_n > 1) = 0$. ♢

All the notions here introduced can be generalized to the case of an n-tuple $(X_1, \dots, X_n)$ of r.v.'s defined on the same sample space. For instance,

Proposition 12.56 *Let $(X_1, \dots, X_n)$ be a vector of n r.v.'s then*

(i) $$E(X_1 + \cdots + X_n) = E(X_1) + \cdots + E(X_n)\ ,$$

(ii) $$var(X_1 + \cdots + X_n) = var(X_1) + \cdots + var(X_n) + 2 \sum_{1\leq i<j\leq n} \Gamma(X_i, X_j)\ .$$

If, moreover, the n r.v.'s $(X_1, \dots, X_n)$ are independent (such that the distribution of $(X_1, \dots, X_n)$ is the product of the distributions of the X_is, $i = 1, \dots, n$, i.e. $P(X_1 = x_1, \dots, X_n = x_n) = P(X_1 = x_1) \cdots P(X_n = x_n)$), then

(iii) $$E(X_1 \cdots X_n) = E(X_1) \cdots E(X_n)\ ,$$

(iv) $$var(X_1 + \cdots + X_n) = var(X_1) + \cdots + var(X_n)\ .$$

Equalities (ii) *and* (iv) *are true provided that the variances $var(X_1)$,..., $var(X_n)$ are finite.*

Proof. (i) is straightforward; (iv) is a consequence of (ii); (iii) is easily proved by induction on n. Let us check (ii); we have

$$\begin{aligned} X_1 + \cdots + X_n - E(X_1 + \cdots + X_n) &= X_1 + \cdots + X_n \\ &\qquad - E(X_1) - \cdots - E(X_n) \\ &= \sum_{i=1}^{n} X_i - E(X_i)\ . \end{aligned}$$

Hence,

$$\begin{aligned} &\big(X_1 + \cdots + X_n - E(X_1 + \cdots + X_n)\big)^2 \\ &\qquad = \sum_{i=1}^{n} \big(X_i - E(X_i)\big)^2 + 2 \sum_{1\leq i<j\leq n} \big(X_i - E(X_i)\big)\big(X_j - E(X_j)\big)\ . \end{aligned}$$

Hence the result is proved by computing the mean. □

12.4.3 Application to approximations

We will give results enabling us to bound the error in approximating an r.v. by its mean.

Theorem 12.57 (Markov's inequality) *Let* $X : \Omega \longrightarrow D \subseteq \mathbb{R}^+$ *be a non-negative r.v., having mean* $E(X) \neq 0$, *and let* λ *be a positive real number, then*

$$\forall \lambda > 0\,, \quad P[X \geq \lambda E(X)] \leq \frac{1}{\lambda}\,. \tag{12.4}$$

Proof. If $0 < \lambda \leq 1$, (12.4) is trivially true, since $\forall A$, $P(A) \leq 1$. Assume $\lambda > 1$; since $X \geq 0$, we have

$$E(X) = \sum_{x \geq 0} xP(X = x) \geq \sum_{x \geq \lambda E(X)} xP(X = x)\,,$$

hence

$$E(X) \;\geq\; \lambda E(X) \sum_{x \geq \lambda E(X)} P(X = x) \;=\; \lambda E(X) P(X \geq \lambda E(X))\,. \qquad \square$$

Theorem 12.58 (Chebyshev's inequality) *Let* X *be an r.v. such that* $E(X)$ *and* $var(X)$ *are both defined; then, for* $\lambda > 0$,

$$P(|X - E(X)| \geq \lambda) \;\leq\; \frac{1}{\lambda^2} var(X). \tag{12.5}$$

Proof. Apply Markov's inequality to the r.v. $Y = (X - E(X))^2$, with $\lambda' = \dfrac{\lambda^2}{var(X)}$; as $E(Y) = var(X)$, we obtain (12.5). □

Exercise 12.23 The average height of a population is 1.65 m and the standard deviation is 0.04 m. Find an upper bound of the probability that the height of a randomly chosen individual in this population is greater than or equal to 1.80 m. ♢

Exercise 12.24 Consider a coin-tossing game with an ideal coin, i.e. such that $P(\text{Tails}) = P(\text{Heads}) = 1/2$. If Tails turns up, the player wins \$1, if Heads turns up, the player loses \$1. Let S_n be the average algebraic 'winnings' after n tosses of the coin. Determine an integer n such that the average winnings S_n are larger than $-1/2$ with a probability greater than or equal to $\dfrac{99}{100}$. Note that, if X_i represents the algebraic 'winnings' at the ith toss, $X_1, \ldots, X_n$ are n independent r.v.'s with the common distribution

$$P(X_i = 1) = P(X_i = -1) = \frac{1}{2}\,.$$

S_n is then defined by

$$S_n = \frac{X_1 + \cdots + X_n}{n}\,. \qquad \diamondsuit$$

We state below the weak law of large numbers. n successive repetitions of a trial can be translated into a sequence X_n of independent r.v.'s with a common distribution. Then, the average (in the arithmetical sense) of the values assumed by the considered variables is likely to lie near the average (in the probabilistic sense), i.e. the mean $E(X)$, as n tends to infinity. We can use the weak law of large numbers to determine the mean of X up to ε, by substituting it with the arithmetical average of the X_n for n large. This justifies *a posteriori* the definition of the mean $E(X)$ of an r.v.

Theorem 12.59 (Weak law of large numbers) *Let $(X_n)_{n\in\mathbb{N}}$ be mutually independent r.v.'s, with a common distribution of mean E and of variance σ^2, let $S_n = X_1 + \cdots + X_n$, and $\varepsilon > 0$. Then*

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - E\right| \geq \varepsilon\right) = 0\,.$$

Proof. We have $E\left(\frac{S_n}{n}\right) = \frac{nE(X_1)}{n} = E$, and $var\left(\frac{S_n}{n}\right) = \sum_{i=1}^{n} var\left(\frac{X_i}{n}\right)$ since the r.v.'s are independent; let

$$var\left(\frac{S_n}{n}\right) = n\frac{\sigma^2}{n^2} = \frac{\sigma^2}{n}\,.$$

Applying Chebyshev's inequality we deduce

$$P\left(\left|\frac{S_n}{n} - E\right| \geq \varepsilon\right) \;\leq\; \frac{\sigma^2}{n\varepsilon^2}\,. \qquad \square$$

The above proof in fact gives a slightly more precise result. If we are able to find an upper bound of σ^2, then we will be able to determine the minimum number n of experiments needed in order to substitute the arithmetic average $\frac{S_n}{n}$ for the mean E with an error less than ε.

12.5 Generating functions

As we represented a sequence by a series (Chapter 8), we can represent an integral-valued r.v. X by a series, since such an r.v. is characterized, for instance, by the sequence $\big(P(X=n)\big)_{n\in\mathbb{N}}$. The advantage is that of giving us a global representation of the r.v. together with its probability distribution and of making the computations we may have to perform much easier.

Definition 12.60 *Let X be an integral-valued r.v. $X\colon\Omega\longrightarrow\mathbb{N}$. Letting $p_n = P(X=n)$, define the generating function g_X of X by the series*

$$g_X(z) \;= \sum_{n=0}^{\infty} p_n z^n = \sum_{n=0}^{\infty} P(X=n) z^n \;=\; E(z^X)\,,$$

g_X will be denoted by g when there can be no ambiguity on X.

The last equality of this definition is a straightforward consequence of Proposition 12.48. Indeed, z^X is a new discrete r.v. obtained by composing X with the function $f\colon \mathbb{N}\longrightarrow\mathbb{R}$ defined by $f(n)=z^n$.

Generating functions thus consist of generating series with non-negative coefficients, and such that $g_X(1) \;= \sum_{n=0}^{\infty} p_n = 1$.

The generating function of X enables us to characterize the mean and the variance of X in a simple way.

Proposition 12.61 *Let X be an integral-valued r.v. with generating function g, then*

(i) $E(X) = g'(1)$,
(ii) $var(X) = g''(1) + g'(1) - (g'(1))^2$,

where g' (resp. g'') is the derivative (resp. the second derivative of g).

Proof.
(i) We have $E(X) = \sum_{n\geq 0} np_n$, or $g'(z) = \sum_{n\geq 0} np_n z^{n-1}$.
(ii) Similarly, we have

$$\begin{aligned} var(X) = E(X^2) - (E(X))^2 &= \sum_{n\geq 1} n^2 p_n - (E(X))^2 \\ &= \sum_{n\geq 1} n^2 p_n - (g'(1))^2. \end{aligned}$$

Note that

$$g''(1) = \sum_{n\geq 2} n(n-1)p_n \quad \text{and} \quad g'(1) = \sum_{n\geq 1} np_n$$

and deduce

$$\sum_{n\geq 1} n^2 p_n = g''(1) + g'(1)\,. \qquad \square$$

The preceding proposition usually gives the simplest way of computing the mean and the variance of an integral-valued r.v..

Exercise 12.25 The Dirichlet generating function of a probability distribution is defined by

$$d(z) = \sum_{n\geq 1} \frac{p_n}{z^n}\,.$$

We thus have $d(1) = 1$. Let X be an r.v. such that $P(X = n) = p_n$; compute $E(X)$, $var(X)$ and $E(\log X)$ in terms of $d(z)$ and of its derivatives. ♢

Proposition 12.62 *Let X and Y be two independent integral-valued r.v.'s, we have $g_{X+Y}(z) = g_X(z)g_Y(z)$, where the product of the generating functions is the product of convolution (or Cauchy product) defined for the generating series.*

Proof. $g_{X+Y}(z) = E(z^{X+Y}) = E(z^X z^Y)$. We check that, if X and Y are two independent r.v.'s, then z^X and z^Y are also independent, since

$$\begin{aligned} P(z^X = x, z^Y = y) &= P(X = \log_z x, Y = \log_z y) \\ &= P(X = \log_z x)P(Y = \log_z y) \\ &= P(z^X = x)P(z^Y = y)\,. \end{aligned}$$

then, by Proposition 12.56, $E(z^X z^Y) = E(z^X)E(z^Y)$, hence

$$g_{X+Y}(z) = g_X(z)g_Y(z). \qquad \square$$

Proposition 12.63 *Let X be an integral-valued r.v. and $a \in \mathbb{N}$, then $g_{aX}(z) = g_X(z^a)$.*

Proof. $g_{aX}(z) = E(z^{aX}) = E((z^a)^X) = g_X(z^a)$. □

Exercise 12.26 Directly prove the preceding result by explicitly computing the generating functions as a series. ♢

Exercise 12.27 Consider a coin-tossing game with probability p for Tails and probability $q = 1 - p$ for Heads. The r.v. S_1 represents the number of tosses required before the first Tails turns up, and the r.v. S_r represents the number of tosses required before the rth Tails turns up.

1. Compute the distribution of S_1 and the generating function of S_1.
2. Compute the generating function of S_r. Note that $S_r = X_1 + \cdots + X_r$, where the X_is are mutually independent and have the distribution of S_1 as common distribution.
3. Deduce the distribution of S_r, its mean and its variance. ♢

EXERCISE 12.28 Let $X_1, \ldots, X_n$ be independent r.v.'s with a common generating function g. Let U be an integral-valued r.v. independent of $X_1, \ldots, X_n$, with generating function f and assuming values in $1, \ldots, n$. Let V be the r.v. $V = \sum_{i=1}^{U} X_i$.

1. Show that $P(V = k) = \sum_{j=1}^{n} P(U = j) \times P((\sum_{i=1}^{j} X_i) = k)$.
2. Compute the generating function of the r.v. V in terms of f and g.
3. Compute the mean and the variance of V in terms of the mean and the variance of U and X_i. ♢

12.6 Common probability distributions

In the present section we give the most usual discrete probability distributions, together with their intuitive motivation, and the main results (mean, variance) will be stated. Explicit computations will be left as exercises; the reader is also advised to have a look at the many excellent handbooks of probability theory (e.g. he/she should benefit by reading Feller, Vol. 1).

12.6.1 Bernoulli trials

These consist of the distribution of a coin-tossing game, with $p = P(\text{Tails})$, $q = 1 - p = P(\text{Heads})$. The corresponding r.v. is $X \colon (\mathbb{B}, \mathcal{P}(\mathbb{B})) \longrightarrow \mathbb{B}$. If we identify Tails with 1 and Heads with 0 then X is defined by: $P(X = 1) = p$ and $P(X = 0) = q$. Its generating function is $g(z) = pz + q$, $E(X) = p$, $var(X) = pq$.

Notation: The Bernoulli distribution is denoted $B(p)$ and p is called the parameter.

12.6.2 Binomial distribution

n independent Bernoulli trials, with the common distribution $B(p)$, are repeated. We are interested in the total number k of 'Tails' produced (we also say 1, or success, for 'Tails', and 0, or failure, for 'Heads'); k is given by the r.v. $S_n = X_1 + \cdots + X_n$, where the X_is are n Bernoulli r.v.'s with the same parameter p. We thus have $g_{S_n}(z) = g_X(z)^n = (pz + q)^n = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} z^k$, by the binomial theorem; hence $P(S_n = k) = \binom{n}{k} p^k q^{n-k}$, $E(S_n) = np$, $var(S_n) = npq$. The binomial distribution is denoted by $b(p, n)$.

EXERCISE 12.29 Let X and Y be two independent r.v.'s with binomial distributions $b(p, m)$ and $b(p, n)$ of respective parameters (p, m) and (p, n). Let $S = X + Y$.

1. What is the distribution of S?
2. Let s be a possible value for S; the conditional distribution of X given that $S = s$ is defined by $P(X = x \,/\, S = s)$, when x ranges over the possible values for X. Find the conditional distribution of X given S. ♢

The binomial distribution is also obtained in sampling problems. Consider a size N population, partitioned into n_1 people of type 1 and n_0 people of type 0, with

$$\frac{n_1}{N} = p, \quad \frac{n_0}{N} = q = 1 - p\,.$$

We randomly choose, n successive times, a person from the entire population (*sampling with replacement*); if S is the number of people drawn of type 1, the distribution of S is a binomial distribution, i.e.

$$P(S = k) = b(p, n)(k) = \binom{n}{k} p^k q^{n-k}\,.$$

Exercise 12.30 Check by a direct combinatorial computation that

$$P(S = k) = b(p, n)(k). \qquad \diamondsuit$$

We can generalize the case of samplings in a population of N people partitioned into

- n_1 individuals of type 1,
- n_2 individuals of type 2,
- ...
- n_r individuals of type r,

with $n_1 + \cdots + n_r = N$ and $\forall i = 1, \ldots, r$: $\dfrac{n_i}{N} = p_i$. Choose a sample with replacement of n individuals; let X_i for $i = 1, \ldots, n$, be the r.v. representing the type of the individual obtained at the ith drawing. We have $P(X_i = j) = p_j$. If S_i, $i = 1, \ldots, r$, is the r.v. representing the number of individuals of type i obtained in the sample, we have:

$$P(S_1 = k_1, \ldots, S_r = k_r) = \frac{n!}{k_1! \cdots k_r!} p_1^{k_1} \cdots p_r^{k_r}\,.$$

Indeed, each sequence $(x_1, \ldots, x_n)$ contains k_i elements of type i, with $k_1 + \cdots + k_r = n$, and it has $p_1^{k_1} \cdots p_r^{k_r}$ probability of occurring. Moreover, there are $n!$ ways of permuting $(x_1, \ldots, x_n)$, but among those $n!$ ways, the $k_i!$ permutations of the k_i elements of type i give the same result, for $i = 1, \ldots, r$.

For $r = 2$, we again find the binomial distribution. We verify that $\forall i$, S_i has a binomial distribution $b(p_i, n)$.

The distribution of $(S_1, \ldots, S_k)$ is called the *multinomial distribution* with parameters $p_1, \ldots, p_k$.

Exercise 12.31

1. Check that S_i has a binomial distribution. For $r = 2$, what is the co-variance $\Gamma(S_1, S_2)$?
2. Generalize the generating functions in order to be able to represent the distribution of $(S_1, \ldots, S_r)$. (Hard.) $\diamondsuit$

12.6.3 Geometric distribution

Let $(X_k)_{k\geq 1}$ be a sequence of independent r.v.'s with a Bernoulli distribution, and X the r.v. representing the number of tosses necessary for the first 'Tails' to turn up. Then, for all $k \geq 1$

$$P(X=k) = P(X_1 = \cdots = X_{k-1} = 1, X_k = 0) = p^{k-1}q\,.$$

We say that X has a *geometric distribution* (or *Pascal distribution*). We have the generating function

$$g(z) = \frac{1-p}{1-zp} = \frac{q}{1-zp}\,,$$

hence

$$E(X) = \frac{p}{q}, \quad var(X) = \frac{p}{q^2}\,.$$

Exercise 12.32 What is the sample space on which X is defined? ♢

Exercise 12.33 A coin is tossed till two successive identical outcomes appear.

1. What is the probability that n tosses are necessary?
2. What is the probability that the experiments stop before the sixth toss?
3. What is the probability that an even number of tosses is necessary? ♢

12.6.4 Hypergeometric distribution

This can be obtained with sampling problems (*sampling without replacement*). Consider, as in 2, a size N population, consisting of n_i individuals of type i, $i = 1,\ldots,r$, $n_1 + \cdots + n_r = N$, $\dfrac{n_i}{N} = p_i$.

Choose a subset of n individuals from the population (sampling without replacement) at the same time. Let S_i be the number of type i individuals among those chosen, $i = 1,\ldots,r$. We verify that

$$P(S_1 = k_1,\ldots,S_r = k_r) = \frac{\binom{n_1}{k_1}\cdots\binom{n_r}{k_r}}{\binom{N}{n}}\,.$$

for $r = 2$, we obtain the *hypergeometric distribution* denoted by $H(N, n_1, n)$, defined by

$$P(S_1 = k) = \frac{\binom{n_1}{k}\binom{N-n_1}{n-k}}{\binom{N}{n}}\,,$$

with $E(S_1) = np_1$, $var(S_1) = np_1(1-p_1)\Big(\dfrac{N-n}{N-1}\Big) = np_1p_2\Big(\dfrac{N-n}{N-1}\Big)$.
Note that in the case of a sampling *with* replacement, we obtain the same mean, but a slightly larger variance, namely, np_1p_2, see Section 12.6.2.

Exercise 12.34
1. Compute $E(S_1)$ and $var(S_1)$ when $r = 2$. Let X_i for $i = 1, \dots, n$, be the r.v. assuming values in $\{1, 2\}$ and representing the type of the individual obtained at the ith drawing. We have $S_1 = \chi_1 + \cdots + \chi_n$, where

$$\chi_i = \chi_{(X_i=1)} = \begin{cases} 1 & \text{if the person chosen at the } i\text{th drawing is of type 1,} \\ 0 & \text{otherwise.} \end{cases}$$

2. Now let r be arbitrary; show that $\forall j = 1, \dots, r$, S_j has a distribution $H(N, n_j, n)$. Deduce $E(S_j)$ and $var(S_j)$. ♢

Asymptotic behaviour

(a) For a fixed n, if $N \to \infty$ and $\dfrac{n_1}{N} \to p$, then $H(N, n_1, n) \to b(p, n)$. Indeed,

$$\begin{aligned} P(S_1 = k) &= \frac{\binom{n_1}{k}\binom{N-n_1}{n-k}}{\binom{N}{n}} = \binom{n}{k}\frac{\binom{N-n}{n_1-k}}{\binom{N}{n_1}} \\ &= \binom{n}{k} n_1(n_1-1)\cdots(n_1-k+1) \\ &\qquad \times \frac{(N-n_1)(N-n_1-1)\cdots(N-n_1-n+k+1)}{N(N-1)\cdots(N-n+1)} \, . \end{aligned}$$

Let

$$p = \frac{n_1}{N} \quad , \quad q = 1 - p \, ,$$

$$\begin{aligned} P(S_1 = k) &= \binom{n}{k} Np(Np-1)\cdots(Np-k+1) \\ &\qquad \times \frac{Nq(Nq-1)\cdots(Nq-n+k+1)}{N(N-1)\cdots(N-n+1)} \\ &\sim \binom{n}{k} p^k (1-p)^{n-k} \qquad \text{when } \ n \to \infty. \end{aligned}$$

The intuition is as follows: for a fixed n, if $N \to \infty$, a drawing without replacement of n individuals in a very large population is close to a drawing with replacement; thus, on these asymptotic conditions, the hypergeometric distribution is close to the binomial distribution.

(b) If n, n_1, N go to infinity and $n\dfrac{n_1}{N} \to \lambda$, then

$$\lim_{n \to \infty} P(S = k) = e^{-\lambda}\frac{\lambda^k}{k!} \, .$$

Exercise 12.35 Verify this result. ♢

12.6.5 Poisson distribution

This is the distribution we have just obtained. An r.v. $X : \Omega \longrightarrow \mathbb{N}$ has a *Poisson distribution* with mean λ, denoted by $p(\lambda)$, if $P(X = k) = e^{-\lambda}\dfrac{\lambda^k}{k!}$, $\forall k \in \mathbb{N}$, with $\lambda > 0$. The generating function of X is given by:

$$g(z) = \textstyle\sum_{n\geq 0} e^{-\lambda} z^n \dfrac{\lambda^n}{n!} = e^{\lambda(z-1)}.$$

Hence we will deduce $E(X) = \lambda$, $var(X) = \lambda$.

Proposition 12.64 *Let X and Y be two independent r.v.'s with Poisson distributions with parameters λ and μ; then $X + Y$ has a Poisson distribution with parameter $\lambda + \mu$.*

Proof. Straightforward, since the generating function g_{X+Y} of $X + Y$ is $g_{X+Y} = g_X g_Y$. □

Proposition 12.65 (Poisson approximation) *Let S_n be an r.v. with a binomial distribution $b(p_n, n)$. Assume that n goes to infinity, with* $\lim_{n\to\infty} np_n = \lambda$, $0 < \lambda < 1$, *then* $\lim_{n\to\infty} P(S_n = k) = e^{-\lambda}\dfrac{\lambda^k}{k!}$.

Proof. Letting $q_n = 1 - p_n$, check the result by induction on k.

- Basis: $k = 0$; then, if $n \to \infty$,

$$P(S_n = 0) = q_n^n = (1 - p_n)^n = \left(1 - \frac{\lambda}{n} + \varepsilon\left(\frac{\lambda}{n}\right)\right)^n \longrightarrow e^{-\lambda}.$$

- Inductive step: assume that $P(S_n = k) \to e^{-\lambda}\dfrac{\lambda^k}{k!}$ for $n \to \infty$.

$$\frac{P(S_n = k+1)}{P(S_n = k)} = \frac{n-k}{k+1} \times \frac{p_n}{q_n} \to \frac{\lambda}{k+1} \qquad \text{when } \ n \to \infty.$$

(Note that $p_n \to 0$ and $q_n \to 1$ for $n \to \infty$.) Thus

$$P(S_n = k+1) \to e^{-\lambda}\frac{\lambda^{k+1}}{(k+1)!} \qquad \text{when } \ n \to \infty.$$ □

Corollary 12.66 *For n 'large' and p 'small' a Poisson distribution with parameter np approximates the binomial distribution $b(p, n)$.*

12.6.6 Uniform distribution

Recall for the sake of completeness the *uniform distribution*: on a finite subset A of $\mathbb{N}$ it is defined by the uniform probability on A:

$$P(X = n) = \begin{cases} \dfrac{1}{|A|} & \text{if } n \in A, \\ 0 & \text{otherwise.} \end{cases}$$

Its generating function is $\sum_{n \in A} \dfrac{z^n}{|A|}$;

$$E(X) = \frac{1}{|A|}(\sum_{n \in A} n) \quad , \quad var(X) = \frac{1}{|A|}(\sum_{n \in A} n^2) - (E(X))^2 \, .$$

Exercise 12.36 Let X and Y be two r.v.'s such that the joint distribution of (X, Y) is given, for all $(m, n) \in \mathbb{N}^2$, such that $m \geq n$, by

$$P(X = n, Y = m) = \frac{\lambda^m}{n!(m-n)!} e^{-2\lambda}.$$

1. Show that the joint distribution of (X, Y) is completely determined.
2. Find the probability distributions of X and of Y. Are X and Y independent?
3. Find the probability distribution of $Y - X$ and the joint distribution of $(X, Y - X)$. Show that the r.v.'s X and $Y - X$ are independent.
4. What are the co-variance $\Gamma(X, Y)$ and the correlation coefficient of $\rho(X, Y)$? ♢

Exercise 12.37 A telephone switchboard receives N daily calls. The r.v. N is assumed to have a Poisson distribution with mean λ. Among these N calls, there are Z wrong numbers, and each number among the N numbers has the probability p of being wrong.

1. Find the probability distribution of (Z, N).
2. Find the probability distribution of Z.
3. Find the probability distribution of N conditioned by Z.
4. Compute the correlation coefficient of $\rho(N, Z)$. ♢