

Moulay Ismail University

École Normale Supérieure (ENS-Meknès)

Department of Sciences

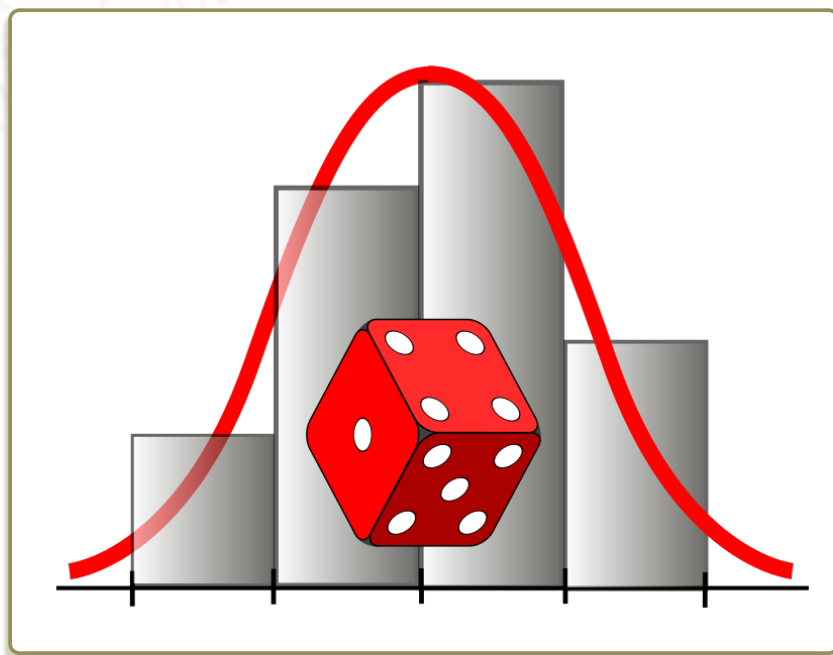
Academic Year : 2025–2026

Program : *Master in Applied Mathematics*

and Educational Engineering (MAIE)

Module Coordinator : *Prof. Youssef El Haoui*

Probabilities, Inferential Statistics, and Their Applications



Elements of Probability

I - Probability on a Finite Set

1) A bit of vocabulary

From a historical point of view, the concept of probability arose from simple examples borrowed from games of chance (the word “hazard/chance” comes from the Arabic *az-zahr* : the die). We introduce this notion through a familiar example : the game of a die.

DEFINITIONS	EXAMPLES
A random experiment is an experiment for which the outcome cannot be predicted in advance. We know the set of all possible outcomes, but we do not know which one will occur. This contrasts with a deterministic experiment, whose result is certain.	Throwing a fair six-sided die is a random experiment. The outcome is the number shown on the upper face of the die.
Every random experiment is associated with a set called the sample space (or fundamental set of the experiment), denoted Ω , containing all possible outcomes.	$\Omega = \{1, 2, 3, 4, 5, 6\}$.
A (random) event is any subset of Ω .	The event “obtaining an even number” is $A = \{2, 4, 6\}$.
An event A is said to be realized (or to <i>occur</i>) if the outcome of the experiment belongs to A .	If the upper face shows 5, then A does not occur. If it shows 4, then A occurs.
An event containing a single outcome is called an elementary event .	$B = \{1\}$ is one of the six elementary events of Ω .

When carrying out a random experiment, the central question we ask is : **How likely is a given event to occur ?**

2) From frequency to the probability of a random event

The theoretical frequency of an event is defined as the limit of its empirical frequency when the number of repetitions of the same experiment tends to infinity (this is a formulation of the Weak Law of Large Numbers).

For example, to estimate the theoretical frequency of obtaining the outcome 3 when throwing a fair die, one may repeat the experiment a large number of times, say 10,000 throws. The theoretical frequency we are looking for, namely, the probability of the elementary event $\{3\}$, will be very close to the empirical frequency observed over these 10,000 trials, and even closer if we increase the number of repetitions to 100,000, and so on. Thus, performing the same experiment many times makes the notion of “chance” much less mysterious.

The formal definition of probability (theoretical frequency) relies on what are known as the **Kolmogorov axioms**.

Three Axioms : Kolmogorov's Axioms

1 Nonnegativity. The probability of any event is always nonnegative :

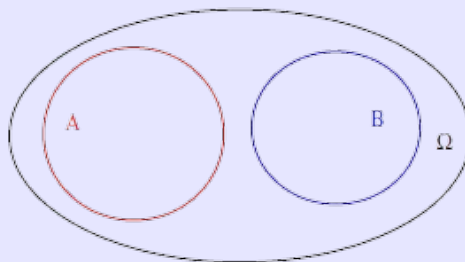
$$p(A) \geq 0.$$

2 Normalization. Since the sample space Ω represents the certain event, it must have maximal probability :

$$p(\Omega) = 1.$$

3 (Finite) additivity. If A and B are *mutually exclusive* (disjoint), then the probability of their union is the sum of their probabilities :

$$p(A \cup B) = p(A) + p(B).$$



II - Probability as a normalized measure

Probability theory, in its modern formulation, is built on the framework of measure theory. A probability is therefore not a new kind of mathematical object : it is simply a *measure* endowed with an additional normalization condition.

Probability as a measure

A probability measure on a measurable space (Ω, \mathcal{F}) is a measure p satisfying

$$p(\Omega) = 1.$$

Here \mathcal{F} denotes a collection of subsets of Ω (the σ -algebra of events). In the finite case treated in this chapter, one usually takes $\mathcal{F} = \mathcal{P}(\Omega)$, the set of all subsets of Ω .

In measure theory, a (general) measure μ on (Ω, \mathcal{F}) is a function

$$\mu : \mathcal{F} \rightarrow [0, +\infty]$$

satisfying :

1 Nonnegativity :

$$\mu(A) \geq 0 \quad \text{for all } A \in \mathcal{F}.$$

2 Null empty set :

$$\mu(\emptyset) = 0.$$

3 Countable additivity : If A_1, A_2, \dots are pairwise disjoint,

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i).$$

A probability measure p satisfies these *same* three axioms. Thus mathematically :

Probability is a measure

For a general measure μ , the total mass $\mu(\Omega)$ can be :

- any nonnegative finite number, or
- $+\infty$.

A *probability* is simply a measure that assigns mass 1 to the entire sample space :

$$p(\Omega) = 1.$$

The number 1 represents the “unit of certainty” (100% of all possible outcomes). Each event receives a portion of this mass, indicating how likely it is to occur.

Probability theory is therefore a special case of measure theory :

$\text{Probability measure} = \text{measure } p \text{ such that } p(\Omega) = 1.$

This perspective is fundamental and will later allow the use of the Lebesgue integral and convergence theorems in probability.

1) Properties of probabilities of a random event

Definition 2.1

If the sample space Ω consists of n elementary events $\{e_i\}$, a *probability measure* on Ω assigns n numbers $p_i \in [0, 1]$ (the probabilities of the elementary events) such that

$$\sum_{i=1}^n p_i = 1.$$

If an event A is the disjoint union of k elementary events $\{e_i\}$ with $0 < k < n$, then by definition

$$p(A) = p\left(\bigcup_{i=1}^k \{e_i\}\right) = \sum_{i=1}^k p(e_i) = \sum_{i=1}^k p_i,$$

hence $0 \leq p(A) \leq 1$.

⚠ The concrete meaning of $p(A)$ is : in a random experiment, the closer $p(A)$ is to 1 (resp. to 0), the more (resp. less) likely A is to occur.

Example 1 : **Uniform probability (equiprobability).**

Assume that all elementary outcomes of the finite sample space Ω are **equally likely**. This means that each outcome $\omega_i \in \Omega$ has the same probability

$$p_i = \frac{1}{n},$$

where $n = \text{card } \Omega$.

Let $A \subset \Omega$ be an event containing k favorable outcomes. Since the probability of each outcome is the same, the probability of A is

$$p(A) = \frac{k}{n} = \frac{\text{card } A}{\text{card } \Omega}.$$

Equivalently,

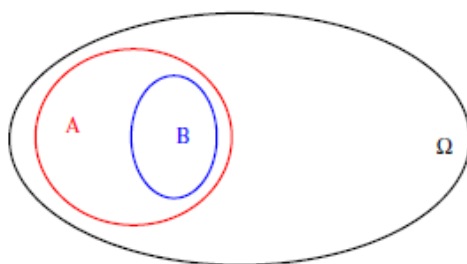
$$p(A) = \frac{\text{number of favorable outcomes}}{\text{number of possible outcomes}}.$$

Property 2.2

If B is a subset of A , then

$$B \subseteq A \Rightarrow p(B) \leq p(A).$$

Indeed, $A = B \cup (A \setminus B)$ with B and $A \setminus B$ disjoint, so $p(A) = p(B) + p(A \setminus B) \geq p(B)$.



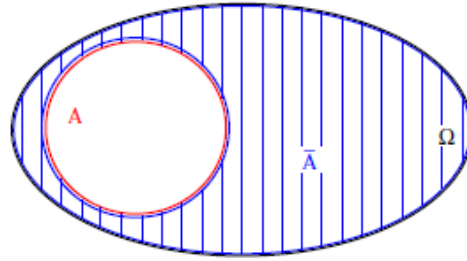
Property 2.3

The event \emptyset is the impossible event ; it never occurs. Its probability is $p(\emptyset) = 0$.

Property 2.4

The *complement* of an event A in Ω is denoted by \bar{A} (or A^c). Its probability is given by

$$p(\bar{A}) = 1 - p(A).$$



Remark 2.5

By Property 2.4 and the non-negativity axiom, we have for every event A :

$$0 \leq p(A) \leq 1.$$

Indeed,

$$p(A) = 1 - p(\bar{A}) \leq 1,$$

since $p(\bar{A}) \geq 0$.

Property 2.6

For any $A, B \subseteq \Omega$,

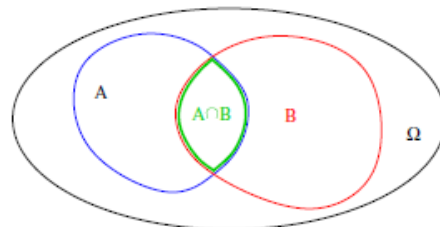
$$p(A \cup B) = p(A) + p(B) - p(A \cap B).$$

Proof:

$p(A) = p(A \setminus B) + p(A \cap B)$ and $p(B) = p(B \setminus A) + p(A \cap B)$ (disjoint unions).

Also $p(A \cup B) = p(A \setminus B) + p(B \setminus A) + p(A \cap B)$.

Adding yields $p(A \cup B) = p(A) + p(B) - p(A \cap B)$. ■

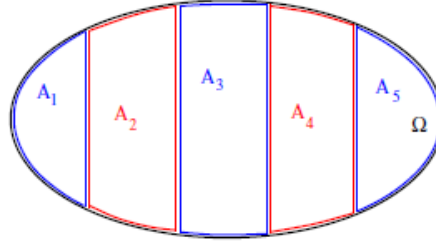


Property 2.7: Addition rule

If A_1, \dots, A_k form a *partition* of Ω (pairwise disjoint and $\Omega = \bigcup_{j=1}^k A_j$), then

$$\sum_{j=1}^k p(A_j) = 1.$$

In this case the A_j are sometimes called a *complete system of events*.



III - Joint probability

The probability that two events A and B both occur is called the joint probability of A and B , denoted $p(A \cap B)$ (read “probability of A and B ”). Its computation depends on whether A and B are dependent or independent.

1) Dependent events - conditional probability

If two events are dependent rather than independent, how do we compute the probability that both occur, since the probability of one depends on the other? We need the notion of conditional probability to quantify the dependence.

Definition 3.1

Let A and B be events, with $p(A) > 0$. The *conditional probability* of B given A is

$$p(B|_A) = \frac{p(A \cap B)}{p(A)}.$$

We read $p(B|_A)$ as “ p of B given A ”.

Remark 3.2

Let (Ω, \mathcal{F}, p) be a probability space and let $B \in \mathcal{F}$ with $p(B) > 0$. Define

$$p_B : \mathcal{F} \rightarrow [0, 1], \quad A \mapsto p_B(A) = p(A|_B) = \frac{p(A \cap B)}{p(B)}.$$

Then p_B is itself a probability measure on (Ω, \mathcal{F}) and satisfies all the usual properties.

Proof:

We check the Kolmogorov axioms for p_B .

(1) Nonnegativity. For any $A \in \mathcal{F}$,

$$p_B(A) = \frac{p(A \cap B)}{p(B)}.$$

Since $p(A \cap B) \geq 0$ and $p(B) > 0$, we have $p_B(A) \geq 0$.

(2) Normalization. We compute

$$p_B(\Omega) = p(\Omega|_B) = \frac{p(\Omega \cap B)}{p(B)} = \frac{p(B)}{p(B)} = 1.$$

(3) Finite additivity (for two disjoint events). Let A_1 and A_2 be two disjoint events in \mathcal{F} . Then $A_1 \cap B$ and $A_2 \cap B$ are also disjoint, and

$$(A_1 \cup A_2) \cap B = (A_1 \cap B) \cup (A_2 \cap B).$$

Therefore, by finite additivity of p ,

$$p_B(A_1 \cup A_2) = \frac{p((A_1 \cup A_2) \cap B)}{p(B)} = \frac{p(A_1 \cap B) + p(A_2 \cap B)}{p(B)} = p_B(A_1) + p_B(A_2).$$

Thus p_B satisfies nonnegativity, normalization, and finite additivity, hence p_B is a probability measure on (Ω, \mathcal{F}) . ■

Theorem 3.3: Law of total probability

Let A_1, \dots, A_k be a complete system of events. Then for any event $B \subseteq \Omega$,

$$p(B) = \sum_{j=1}^k p(B \cap A_j) = \sum_{j=1}^k p(A_j) p(B|A_j). \quad (\text{Total probability formula})$$

Proof:

If $i \neq j$, then $(A_i \cap B) \cap (A_j \cap B) = (A_i \cap A_j) \cap B = \emptyset$. By distributivity, $B \cap \Omega = B \cap (A_1 \cup \dots \cup A_k) = (A_1 \cap B) \cup \dots \cup (A_k \cap B)$. By the additivity property (finite additivity for disjoint unions),

$$p(B) = \sum_{j=1}^k p(A_j \cap B).$$

By the definition of conditional probability, $p(A_j \cap B) = p(B|A_j) p(A_j)$, hence

$$p(B) = \sum_{j=1}^k p(B|A_j) p(A_j).$$

■

Application exercise :

Suppose there are three different routes to ENS-Meknès : A, B, and C. The probabilities of taking them are 30%, 40%, and 30% respectively. The probability of arriving late if you take A is 10%, if you take B is 5%, and if you take C is 15%. Find the probability of arriving late to ENS-Meknès to study.

Solution : Let R be “Arrive late”, A_1 “Take route A”, A_2 “Take route B”, A_3 “Take route C”. We know $p(A_1) = 0.3$, $p(A_2) = 0.4$, $p(A_3) = 0.3$, and $p(R|A_1) = 0.1$, $p(R|A_2) = 0.05$, $p(R|A_3) = 0.15$. By total probability,

$$p(R) = \sum_{j=1}^3 p(A_j) p(R|A_j) = 0.03 + 0.02 + 0.045 = \boxed{0.095}.$$

Thus the probability of arriving late is 9.5%.

Theorem 3.4: Bayes' formula

Let A_1, \dots, A_k be a complete system of events and let B be an event with $p(B) > 0$. Then

$$p(A_j|B) = \frac{p(A_j \cap B)}{p(B)} = \frac{p(A_j) p(B|A_j)}{\sum_{i=1}^k p(A_i) p(B|A_i)}. \quad (\text{Bayes' formula})$$

Application of Bayes' formula :

A medical lab produces a screening test for a disease. Sensitivity is 95% (positives among the truly ill). Specificity is 90% (negatives among the healthy). Prevalence in the population is 1%. If a person tests positive, what is the probability they are actually ill?

Solution : Let M = “person is ill”, T = “test is positive”. Given $p(M) = 0.01$, $p(T|M) = 0.95$, $p(T|\bar{M}) = 0.10$. We seek $p(M|T)$:

$$p(M|T) = \frac{p(M)p(T|M)}{p(M)p(T|M) + p(\bar{M})p(T|\bar{M})} = \frac{0.01 \cdot 0.95}{0.01 \cdot 0.95 + 0.99 \cdot 0.10} \approx \boxed{0.087}.$$

So a positive test implies about an 8.7% chance of actually being ill.

Remark 3.5

Sampling *without* replacement is a good illustration of dependent events.

Exercise

An urn contains 5 black balls and 3 white balls. Two balls are drawn without replacement. What is the probability of drawing two white balls?

Solution : Let B_1 be “white on the first draw”, B_2 “white on the second draw”. Then $p(B_1 \cap B_2) = p(B_1) p(B_2|B_1) = \frac{3}{8} \cdot \frac{2}{7} = \boxed{\frac{3}{28}}$.

2) Independent events**Definition 3.6: Independence of two events**

Two events A and B are independent if $p(A|_B) = p(A)$.

Property 3.7

If two events are independent, the probability that both occur equals the product of their probabilities :

$$p(A \cap B) = p(A) p(B). \quad (\text{Product rule for independent events})$$

Example 2 :

For a fair six-sided die, let A : “the result is even”, and B : “the result is a multiple of three”. Then $A = \{2, 4, 6\}$, $B = \{3, 6\}$, $A \cap B = \{6\}$ and $p(A) = 3/6$, $p(B) = 2/6$, $p(A \cap B) = 1/6 = p(A)p(B)$; hence A and B are independent.

Example 3 :

Throw a red die and a green die. What is $p(\text{total} = 2)$? We need a 1 on each die. The events “red die = 1” and “green die = 1” are independent, each of probability $1/6$, so

$$p(\text{total} = 2) = \boxed{\frac{1}{36}}.$$

Remark 3.8

Sampling *with* replacement is a good illustration of independent events.

Definition 3.9: Mutual independence of n events

For $n \geq 2$, events A_1, \dots, A_n are mutually independent (independent as a family) if

$$p(A_1 \cap \dots \cap A_n) = \prod_{i=1}^n p(A_i).$$

Remark 3.10

vents can be pairwise independent ($p(A_i \cap A_j) = p(A_i)p(A_j)$ for $i \neq j$) without being mutually independent.

Example 4 :

Throw two fair dice and consider

A_1 : “the first die is even”,

A_2 : “the second die is even”,

A_3 : “the sum of the two throws is even”.

They are pairwise independent but not mutually independent. Indeed,

$$\text{card}(\Omega) = 36, \quad p(A_1) = p(A_2) = p(A_3) = \frac{1}{2}, \quad p(A_i \cap A_j) = \frac{1}{4} \quad (i \neq j),$$

but $p(A_1 \cap A_2 \cap A_3) = \frac{1}{4} \neq \frac{1}{8} = p(A_1)p(A_2)p(A_3)$.

IV - How to tackle a probability exercise ?

In many situations, problem solving is helped by the following approach.

1 List the elementary events or describe the sample space Ω .

2 Determine the probability measure on Ω .

- If the probability is uniform, then $p(A) = \frac{\text{card } A}{\text{card } \Omega}$.
- Otherwise, determine each elementary probability, remembering that their sum is 1.

3 Identify precisely the event(s) whose probability is sought.

4 Use the appropriate formula : Are we computing the probability of

- an elementary event ?
- a complementary event ?
- compatible/incompatible events (total probability) ?
- dependent/independent events (product rule) ?

Exercise

1. Two fair dice are thrown. What is the probability of getting a total of 7 ?
2. Now the dice are loaded : even numbers are twice as likely as odd numbers. What is $p(\text{sum} \neq 8)$?

Solution.

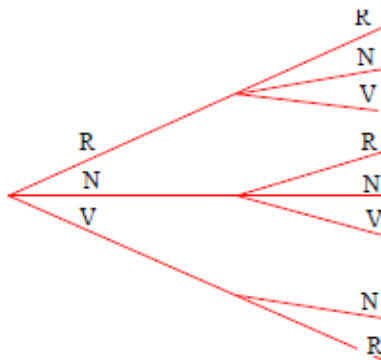
1.
 - $\Omega = \{1, \dots, 6\} \times \{1, \dots, 6\}$, so $\text{card}(\Omega) = 36$.
 - The measure is uniform.
 - The event (sum = 7) has 6 favorable pairs : (1, 6), (6, 1), (2, 5), (5, 2), (3, 4), (4, 3).
 - Hence $p(\text{sum} = 7) = \frac{6}{36} = \boxed{\frac{1}{6}}$.
2.
 - Same Ω .
 - Let p_p be the probability of an even face and p_i that of an odd face ; the statement gives $p_p = 2p_i$. Since the six outcomes exhaust the space : $3p_p + 3p_i = 1 \Rightarrow 9p_i = 1$, so $p_i = \frac{1}{9}$, $p_p = \frac{2}{9}$.
 - Compute $p(\text{sum} = 8)$ first : elementary outcomes are (2, 6), (6, 2), (3, 5), (5, 3), (4, 4).
 - Using independence across dice, $p\{(2, 6)\} = p\{2\}p\{6\} = \frac{2}{9} \cdot \frac{2}{9} = \frac{4}{81}$; likewise $p\{(6, 2)\} = p\{(4, 4)\} = \frac{4}{81}$ and $p\{(5, 3)\} = p\{(3, 5)\} = \frac{1}{81}$.
 - Thus $p(\text{sum} = 8) = \frac{14}{81}$ and $p(\text{sum} \neq 8) = 1 - \frac{14}{81} = \boxed{\frac{67}{81}}$.

V - Counting techniques

1) Trees

Example 5 :

Consider an urn with two red, two black, and one green ball. Draw two balls without replacement. This is a two-stage experiment ; its possibilities are represented by a (horizontal) tree. We get three main branches and three secondary branches at each stage, except when a green ball is drawn first. The number of terminal branches equals the number of outcomes in Ω .



2) Fundamental Principle of Counting

Fundamental Principle of Counting (Multiplication Axiom)

If one experiment can be performed in m ways and a second in n ways, then doing the first followed by the second can be performed in $m \times n$ ways.

This principle is not limited to two experiments ; it applies to any finite number.

Example 6 :

You have 3 shirts (red, blue, green) and 4 pants (black, beige, blue, gray). In how many ways can you choose a shirt followed by pants to make an outfit ? Answer : $3 \times 4 = 12$ outfits.

3) Arrangements and permutations

Consider a set of n distinct objects. Choose p of them and order them.

Definition 5.1

An ordered selection of p distinct objects from n is called an *arrangement* (with $p \leq n$).

How many ? View the p positions as fixed ; the number of choices is n for the first, $n - 1$ for the second, ..., $n - p + 1$ for the p -th.

Proposition 5.2

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

$\triangle!$ $n!$ (“ n factorial”) is the product of all integers up to n : $n! = n(n-1) \cdots 3 \times 2 \times 1$, with the convention $0! = 1$.

Example 7 :

Arrangements of two letters from $\{a, b, c, d\}$: $A_4^2 = 4 \cdot 3 = \boxed{12}$,
namely $(a, b), (a, c), (a, d), (b, a), (b, c), (b, d), (c, a), (c, b), (c, d), (d, a), (d, b), (d, c)$.

Example 8 :

If $\text{card}(A) = s$ and $\text{card}(B) = t$, the number of injective functions $A \rightarrow B$ is A_t^s .
If $t < s$, no injection exists.

Definition 5.3

A *permutation* of n elements is an ordered arrangement of all n elements.

Proposition 5.4

The number of permutations of n elements is $A_n^n = n!$.

4) Combinations

Definition 5.5

A selection of p distinct objects from n *without regard to order* is called a *combination* of p from n .

In the previous example with $\{a, b, c, d\}$, the combination $\{a, b\}$ is the same as $\{b, a\}$, whereas the arrangements (a, b) and (b, a) are different.

How many ? The number of combinations of p from n is denoted by the English notation

$$\binom{n}{p} \quad (\text{read : “}n \text{ choose } p\text{”}).$$

Since each unordered selection corresponds to $p!$ ordered arrangements :

Proposition 5.6

$$\binom{n}{p} = \frac{A_n^p}{p!} = \frac{n(n-1) \cdots (n-p+1)}{p!} = \frac{n!}{p!(n-p)!}.$$

Example 9 : Combinations of two letters from $\{a, b, c, d\}$:

$$\binom{4}{2} = \frac{4!}{2!2!} = \boxed{6},$$

namely $\{a, b\}, \{a, c\}, \{a, d\}, \{b, c\}, \{b, d\}, \{c, d\}$.

Example 10 : At a party of ten people, how many handshakes occur if everyone shakes hands with everyone else ?

$$\binom{10}{2} = 45 \text{ handshakes.}$$

5) Case where elements need not be distinct

How many ordered selections of p elements from n when repetition is allowed ? Each position has n choices, so by the multiplication axiom there are n^p choices (here p may exceed n).

Example 11 : If $E = \{a, b, c, d, e\}$, there are $5^3 = 125$ three-letter words.

6) Summary

	Conditions	Number of possible selections	A common example :
possible with $p \geq n$	p elements not necessarily distinct, ordered	p -lists of elements of $E : n^p$	successive draws <i>with</i> replacement of p objects from n .
$p < n$	p elements all distinct and ordered	arrangements of p elements of $E : A_n^p$	successive draws <i>without</i> replacement.
$p = n$	n elements distinct and ordered	permutations of the n elements : $n!$	anagrams of a word with distinct letters.
$p < n$	p elements distinct and unordered	combinations of p from $n : \binom{n}{p}$	simultaneous draw of p objects from n .

Random Variables

I - Definitions

Example 1 :

We toss a fair coin twice and are interested in the number of times the side “heads” appears. To compute the probabilities of the various outcomes, we introduce a variable X denoting the number of “heads” obtained. The variable X can take the values 0, 1, 2.

Example 2 :

A dart is thrown at a circular target of radius 50 cm, and we are interested in the distance between the dart and the center. We introduce a variable X , defined as the distance between the impact point and the center of the target. This variable can take any real value between 0 and 50.

In both cases, X takes real values that depend on the outcome of the random experiment. The values taken by X are therefore random. Thus, X is called a *random variable*.

Example 3 :

Suppose the sample space Ω is the set of students. We select one student at random (see graphic 2.1). To each student we associate two numerical quantities :

1. their weight in kilograms,
2. their height in meters.

Let W and H be the functions that assign to every student the numerical values w (weight) and h (height), respectively. Both W and H are **random variables**, since they take an element of the sample space (a student) and return a real number.

The random variable W takes as input a student and outputs his weight. Similarly, H takes a student and outputs his height.

From these two random variables, we can construct a third one : the *body mass index* (BMI), defined by

$$B = \frac{W}{H^2}.$$

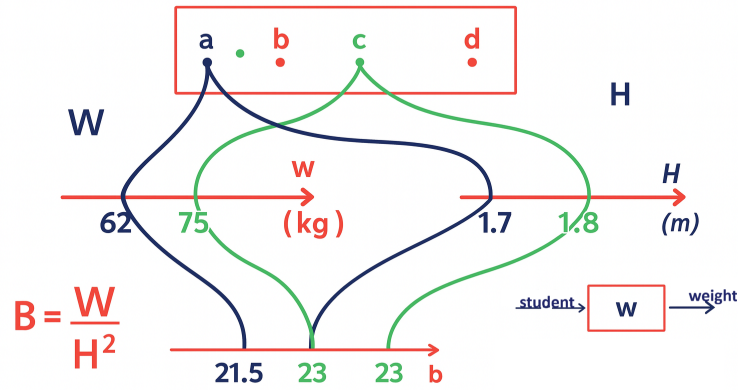


Figure 2.1 – Random variables associated with a randomly chosen student

Definition 1.1

Let Ω be a sample space associated with a random experiment, equipped with a probability measure. A *random variable* (r. v.) is a mapping

$$X : \Omega \longrightarrow \mathbb{R}.$$

- 👉 A random variable is a function that assigns numerical values to random outcomes.
- 👉 By convention, a random variable is denoted by an uppercase letter (e.g., X), while the values it can take are denoted by lowercase letters ($x_1, x_2, \dots, x_i, \dots, x_n$).
- 👉 The two random variables defined in the first two examples are of different types : the first is discrete, the second continuous.
- 👉 We can have several random variables defined on the same sample space.

II - Discrete Random Variables

Definition 2.1: Discrete random variable

A discrete random variable is a random variable that **takes only integer values**, in a finite or countably infinite subset of \mathbb{Z} .

To understand a random variable, it is essential to know which values occur most often and which are less frequent. Concretely, this amounts to determining the probabilities associated with each possible value of the variable.

Definition 2.2: Probability mass function

Associating to each possible value of a random variable the probability that the variable takes this value defines the *probability law* or *probability distribution* of the variable.

Let X be a discrete random variable taking values x_i . To compute $p(X = x_i)$, we collect all elementary events e_j such that $X(e_j) = x_i$, and we obtain

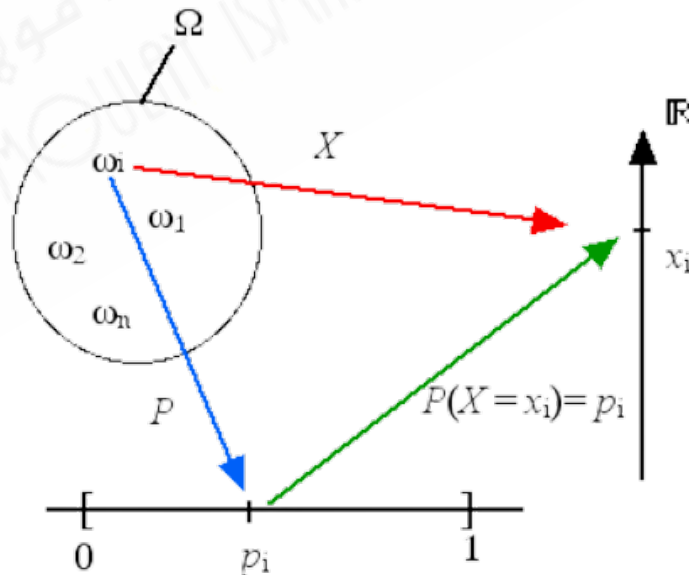
$$p(X = x_i) = p(\{\omega \in \Omega : X(\omega) = x_i\}).$$

The *probability mass function* (pmf) of X is the function

$$f : \mathbb{R} \longrightarrow [0, 1], \quad f(x_i) = p(X = x_i),$$

and, of course,

$$\sum_i f(x_i) = 1.$$



Example 4 : Case of Example 1.
We have

$$\Omega = \left\{ \underbrace{(\text{tails}, \text{tails})}_{e_1}, \underbrace{(\text{tails}, \text{heads})}_{e_2}, \underbrace{(\text{heads}, \text{tails})}_{e_3}, \underbrace{(\text{heads}, \text{heads})}_{e_4} \right\},$$

and the variable

$X = \text{“number of heads”}$

takes the values

$$X(\Omega) = \{0, 1, 2\}.$$

The pmf of X is :

$$f(0) = p(X = 0) = p(e_1) = \frac{1}{4},$$

$$f(1) = p(X = 1) = p(e_2) + p(e_3) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2},$$

$$f(2) = p(X = 2) = p(e_4) = \frac{1}{4},$$

$$f(x) = 0 \quad \text{if } x \notin \{0, 1, 2\}.$$

Its probability distribution can be presented in a table :

Value x_i	0	1	2	Total
$p(X = x_i)$	1/4	1/2	1/4	1

1) Cumulative distribution function

In descriptive statistics (see the course “Probabilities and Descriptive Statistics” in semester 1), the notion of increasing cumulative frequencies is equivalent, in probability theory, to the cumulative distribution function.

Definition 2.3: Cumulative distribution function in the discrete case

The cumulative distribution function of a random variable X gives, for each real value x , the probability that X takes a value less than or equal to x . It is the sum of the probabilities of the values of X up to x . It is denoted F_X , and, when unambiguous, simply F .

$$\forall x \in \mathbb{R} : F_X(x) = p(X \leq x) = \sum_{x_i \leq x} p(X = x_i).$$

The cdf is always increasing, takes values between 0 and 1, and is a very useful tool in theoretical work.

III - Continuous Random Variables

Definition 3.1: Continuous random variable

A random variable is said to be *continuous* if it can take all values of a finite or infinite interval.

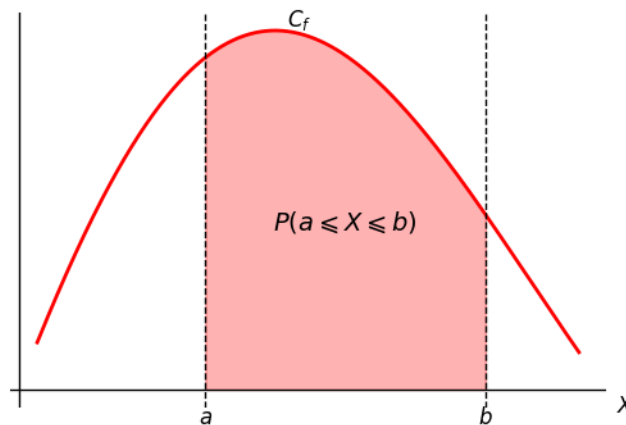
1) Probability density function

Definition 3.2

A function f is a probability density function if it has the following properties :

- ✓ It is nonnegative ;
- ✓ It is integrable over \mathbb{R} ,
such that the probability that the random variable X lies between a and b , i.e., $p(a \leq X \leq b)$, equals the area between the x -axis, the graph of f , and the vertical lines $x = a$ and $x = b$,

$$p(a \leq X \leq b) = \int_a^b f(x) dx$$



- ✓ The total area under the graph is 1 :

$$\int_{\mathbb{R}} f(x) dx = 1.$$

Remark 3.3

Note that for a continuous random variable X we always have

$$p(X = a) = 0.$$

Indeed,

If f is continuous on an interval of the form $[a, a + h]$ with $h \rightarrow 0^+$, then

$$p(a \leq X \leq a + h) = \int_a^{a+h} f(x) dx = h f(a + \theta h) \text{ with } (0 < \theta < 1) \text{ (Mean Value Theorem).}$$

Thus, as $h \rightarrow 0$, $f(a + \theta h) \rightarrow f(a)$ and $h f(a + \theta h) \rightarrow 0$,

whence $p(a \leq X \leq a + h) \rightarrow p(X = a) = 0$.

2) Cumulative distribution function

Definition 3.4: Cdf in the continuous case

As for discrete random variables, one defines the cumulative distribution function F of a continuous random variable X :

$$F(x) = p(X \leq x) = \int_{-\infty}^x f(t) dt.$$

Property 3.5

1. F is continuous on \mathbb{R} , differentiable at every point where f is continuous, and $F' = f$.
2. F is increasing on \mathbb{R} .
3. $\lim_{x \rightarrow -\infty} F(x) = 0$, $\lim_{x \rightarrow +\infty} F(x) = 1$.
4. $p(a \leq X \leq b) = F(b) - F(a)$.

Proof:

1. Follows from $F(x) = \int_{-\infty}^x f(t) dt$.
2. Since $F' = f$, it is nonnegative on \mathbb{R} .
3. $\int_{-\infty}^x f(t) dt \rightarrow 0$ as $x \rightarrow -\infty$.
4. $p(a \leq X \leq b) = p(X < b) - p(X < a) = F(b) - F(a)$, hence $\int_{-\infty}^b f(t) dt - \int_{-\infty}^a f(t) dt = \int_a^b f(t) dt$.

■

Exercise 1

Let f be the function on \mathbb{R} defined by

$$f(x) = \begin{cases} 0 & \text{if } x < 1, \\ \frac{2}{x^3} & \text{otherwise.} \end{cases}$$

1. Show that f is a probability density function for a random variable X .
2. Determine the cumulative distribution function F_X of X .

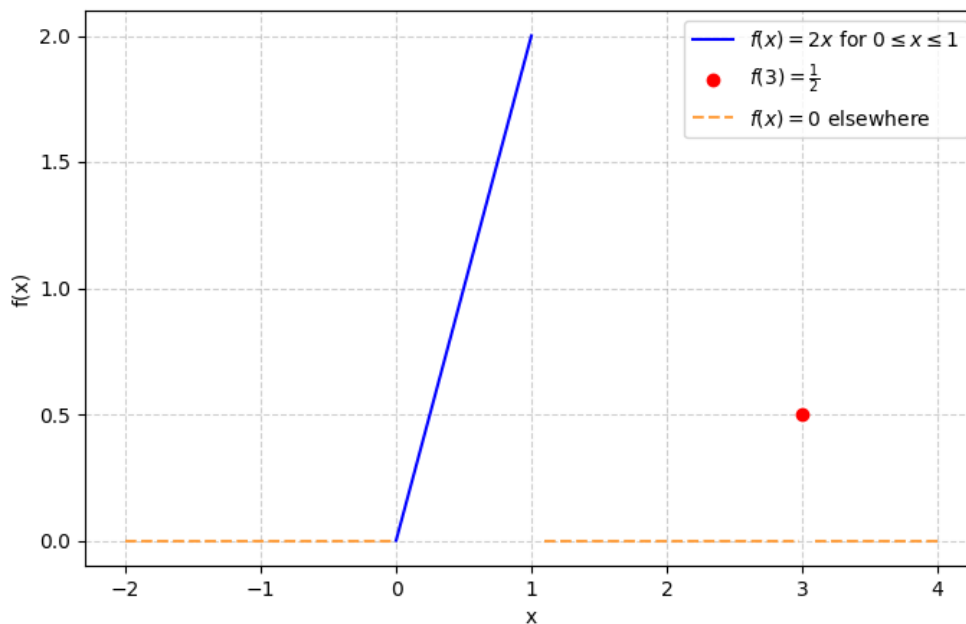
Example 5 :

Let f be the function on \mathbb{R} defined by

$$f(x) = \begin{cases} 2x & \text{if } 0 \leq x < 1, \\ \frac{1}{2} & \text{if } x = 3, \\ 0 & \text{otherwise.} \end{cases}$$

1. Show that f is a probability density function for a random variable X .
2. Determine the cumulative distribution function F_X of X .

Solution :



1. f is clearly nonnegative, and piecewise integrable : on $] - \infty, 0[$, $]0, 1[$, $]0, 3[$ and $]3, +\infty[$ it is locally integrable. The right and left limits at 0, 1, and 3 exist and are finite. Moreover,

$$\int_{\mathbb{R}} f(t) dt = \int_{-\infty}^0 0 dt + \int_0^1 2t dt + \int_1^3 0 dt + \int_3^{+\infty} 0 dt = 1.$$

2.
 • If $x < 0$, $F(x) = \int_{-\infty}^x 0 dt = 0$,
 • If $0 \leq x \leq 1$, $F(x) = \int_{-\infty}^0 0 dt + \int_0^x 2t dt = [t^2]_0^x = x^2$,
 • If $x > 1$, $F(x) = \int_{-\infty}^0 0 dt + \int_0^1 2t dt + \underbrace{\int_1^x f(t) dt}_{=0} = 1$,

$$F(x) = \begin{cases} 0 & \text{if } x < 0, \\ x^2 & \text{if } 0 \leq x \leq 1, \\ 1 & \text{if } x \geq 1. \end{cases}$$

Example 6 :

Let f be the function on \mathbb{R} defined by

$$f(x) = \begin{cases} k \exp(-x) & \text{if } x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

1. Determine k so that f is the probability density function of a random variable X .
2. Determine the cumulative distribution function of X .
3. Compute $p(1 < X < 2)$.

Solution :

1. f must be nonnegative, so k must be positive. Moreover, f is Riemann-integrable on \mathbb{R} (piecewise integrable : on $] - \infty, 0[$ and $]0, +\infty[$ it is locally integrable; the limits $\lim_{0^+} f(x), \lim_{0^-} f(x)$ exist and are finite; and $\lim_{A \rightarrow -\infty} \int_A^0 f(t) dt$ and $\lim_{B \rightarrow +\infty} \int_0^B k e^{-t} dt$ are finite).
Furthermore, a probability density must satisfy $\int_{\mathbb{R}} f(t) dt = 1$,
hence $\int_0^{+\infty} k e^{-t} dt = 1$. It follows that $\boxed{k = 1}$.
2. By definition, the cdf of X is

$$F(x) = \begin{cases} \int_0^x e^{-t} dt = 1 - e^{-x} & \text{if } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

3.

$$p(1 < X < 2) = F(2) - F(1) = \boxed{e^{-1} - e^{-2}} \approx 0.23.$$

IV - Descriptive Parameters of a Probability Distribution

1) Expectation of a probability distribution

Definition 4.1

1. Discrete case.

👉 Let X be a discrete random variable taking finitely many values x_1, x_2, \dots, x_n with pmf $f : f(x_i) = p(X = x_i)$. The *expectation* of X , denoted $E(X)$, is

$$E(X) = \sum_{i=1}^n x_i f(x_i).$$

👉 If X takes a countable set of values $x_1, x_2, \dots, x_n, \dots$, its *expectation* is

$$E(X) = \sum_{i=1}^{\infty} x_i f(x_i),$$

provided the series converges absolutely.

2. Continuous case. If X is continuous with density f , its *expectation* is

$$E(X) = \int_{\mathbb{R}} x f(x) dx,$$

provided the function $x \mapsto x f(x)$ is integrable on \mathbb{R} .

2) Variance of a Probability Distribution

Definition 4.2

1. The *variance* of a random variable X is the mean of the squared deviations from its mean :

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

The computation simplifies using :

König–Huygens formula

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

2. The *standard deviation* of X is the square root of its variance :

$$\sigma(X) = \sqrt{V(X)}.$$

In the case of a finite discrete random variable,

$$V(X) = \sum_{i=1}^n (x_i - E(X))^2 f(x_i) = \left(\sum_{i=1}^n x_i^2 f(x_i) \right) - E(X)^2.$$

In the case of a continuous random variable,

$$V(X) = \int_{\mathbb{R}} (x - E(X))^2 f(x) dx = \left(\int_{\mathbb{R}} x^2 f(x) dx \right) - E(X)^2.$$

3) Properties of expectation and variance

We summarize the main properties of these two parameters in a table.

Shift (origin)	Scale change	Affine transformation
$E(X + c) = E(X) + c$	$E(aX) = aE(X)$	$E(aX + c) = aE(X) + c$
$V(X + c) = V(X)$	$V(aX) = a^2V(X)$	$V(aX + c) = a^2V(X)$
$\sigma(X + c) = \sigma(X)$	$\sigma(aX) = a \sigma(X)$	$\sigma(aX + c) = a \sigma(X)$

Definition 4.3

- A random variable X is *centered* if its expectation is zero.
- A random variable X is *reduced* if its standard deviation equals 1.
- A centered and reduced random variable is said to be *standardized*.

4) A Pair of Random Variables

Marginal laws. From the joint law of (X, Y) , one recovers the marginal laws :

$$p(X = x) = \sum_{y \in Y(\Omega)} p(X = x, Y = y), \quad p(Y = y) = \sum_{x \in X(\Omega)} p(X = x, Y = y).$$

(If X or Y take infinitely many values, the corresponding series is absolutely convergent.)

Expectation. For any function $f : X(\Omega) \times Y(\Omega) \rightarrow \mathbb{R}$,

$$E[f(X, Y)] = \sum_x \sum_y f(x, y) p(X = x, Y = y).$$

In particular,

$$E(XY) = \sum_x \sum_y xy p(X = x, Y = y).$$

This directly yields the linearity :

$$E(X + Y) = E(X) + E(Y),$$

and, if X and Y are independent,

$$E(XY) = E(X) E(Y).$$

Proposition 4.4: Linearity of the Expectation

Let X and Y be real-valued random variables with finite expectations, and let $a, b \in \mathbb{R}$. Then

- 1) $E(aX + b) = a E(X) + b$,
- 2) $E(X + Y) = E(X) + E(Y)$.

Proof:

We give the proof in the discrete case. Let $p_{X,Y}(x, y)$ be the joint pmf, and

$$p_X(x) = \sum_y p_{X,Y}(x, y), \quad p_Y(y) = \sum_x p_{X,Y}(x, y)$$

its marginals.

1) **Affine change of variable.**

$$E(aX + b) = \sum_x (ax + b) p_X(x) = a \sum_x x p_X(x) + b \sum_x p_X(x) = aE(X) + b.$$

2) **Sum of two random variables.**

$$E(X + Y) = \sum_x \sum_y (x + y) p_{X,Y}(x, y) = \sum_x x p_X(x) + \sum_y y p_Y(y) = E(X) + E(Y).$$

■

Remark 4.5

The same properties hold for continuous variables. If (X, Y) has joint density $f_{X,Y}$ and marginals

$$f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x, y) dy, \quad f_Y(y) = \int_{\mathbb{R}} f_{X,Y}(x, y) dx,$$

then, by Fubini's theorem,

$$E(X + Y) = E(X) + E(Y).$$

The identity $E(aX + b) = aE(X) + b$ follows from linearity of integrals.

4- 1) Covariance**Definition 4.6: Covariance**

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y).$$

(For infinite sums, the expectation is well-defined if the corresponding series are absolutely convergent.)

Covariance behaves linearly :

$$\text{Cov}(aX + b, cY + d) = ac \text{Cov}(X, Y).$$

If X and Y are independent, then

$$\text{Cov}(X, Y) = 0.$$



The converse is false.

Proposition 4.7

1. If X and Y are independent, then $\text{Cov}(X, Y) = 0$.
2. In general,

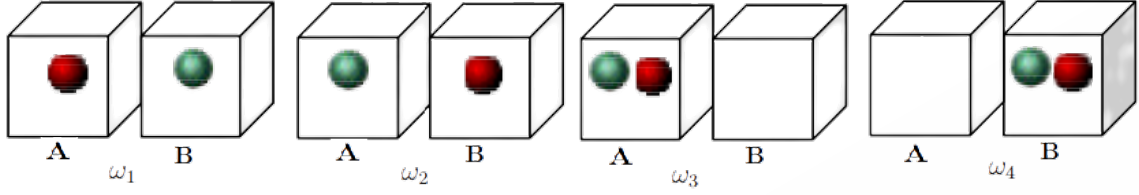
$$E(XY) = E(X)E(Y) + \text{Cov}(X, Y), \quad V(X + Y) = V(X) + V(Y) + 2 \text{Cov}(X, Y).$$

Example 7 :

Two balls, one red and one green, are placed at random into boxes A and B .
Define : X = number of balls in box A ; Y = number of empty boxes.

1) **Sample space.**

$$\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}.$$



2) **Distribution of X .**

$$X(\omega_1) = 1, X(\omega_2) = 1, X(\omega_3) = 2, X(\omega_4) = 0, \quad X(\Omega) = \{0, 1, 2\}.$$

$$p(X = 0) = \frac{1}{4}, \quad p(X = 1) = \frac{1}{2}, \quad p(X = 2) = \frac{1}{4}.$$

$$E(X) = 1, \quad V(X) = \frac{1}{2}.$$

3) **Distribution of Y .**

$$Y(\omega_1) = 0, Y(\omega_2) = 0, Y(\omega_3) = 1, Y(\omega_4) = 1, \quad Y(\Omega) = \{0, 1\}.$$

$$p(Y = 0) = \frac{1}{2}, \quad p(Y = 1) = \frac{1}{2}, \\ E(Y) = \frac{1}{2}, \quad V(Y) = \frac{1}{4}.$$

4) **Distribution of $X + Y$.**

$$X + Y(\Omega) = \{1, 3\}, \\ p(X + Y = 1) = \frac{3}{4}, \quad p(X + Y = 3) = \frac{1}{4}, \\ E(X + Y) = \frac{3}{2}, \quad V(X + Y) = \frac{3}{4}.$$

5) **Distribution of XY .**

$$XY(\Omega) = \{0, 2\}, \\ p(XY = 0) = \frac{3}{4}, \quad p(XY = 2) = \frac{1}{4}, \\ E(XY) = \frac{1}{2}, \quad V(XY) = \frac{3}{4}.$$

6) **Independence.** Although

$$E(XY) = E(X)E(Y),$$

we have

$$p(X = 0, Y = 0) = 0 \neq \frac{1}{4} \cdot \frac{1}{2} = \frac{1}{8}.$$

Hence X and Y are *not independent*.

V - Markov and Chebyshev Inequalities

When studying a random variable X , it is often difficult, or impossible, to know its exact PMF or PDF. Even with limited information (such as its mean or variance), we can still obtain useful bounds using **concentration inequalities**, which estimate how likely X is to take unusually large values or deviate from its mean.

1) Markov Inequality

- ☞ Provides bounds for non-negative random variables.
- ☞ If $E(X)$ is small, then large values of X must be rare.

Proposition 5.1: Markov Inequality

If $X \geq 0$ and $a > 0$, then

$$p(X \geq a) \leq \frac{E(X)}{a}.$$

Proof:

For continuous X :

$$E(X) = \int_0^\infty x f_X(x) dx \geq \int_a^\infty x f_X(x) dx \geq a \int_a^\infty f_X(x) dx = a p(X \geq a).$$

Hence the result. (The discrete case is identical, with sums instead of integrals.) ■

2) Chebyshev Inequality

- ☞ Applies to any random variable with finite mean μ and variance σ^2 .
- ☞ Quantifies how rarely X deviates far from its mean.

Proposition 5.2: Chebyshev Inequality

$$p(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}, \quad c > 0.$$

Proof:

Since $(X - \mu)^2 \geq 0$, applying Markov's inequality to it gives

$$p(|X - \mu| \geq c) = p((X - \mu)^2 \geq c^2) \leq \frac{E((X - \mu)^2)}{c^2} = \frac{\sigma^2}{c^2}.$$

■

Classical Probability Distributions

I - Discrete Distributions

1) The Binomial Distribution

1- 1) Bernoulli Scheme

Let us consider a random experiment represented by a sample space Ω . We choose an event A and call its occurrence a “success,” with probability $p(A)$, denoted p . Its complement, called a “failure,” has probability $q = 1 - p$.

Definition 1.1: Bernoulli Distribution

A random variable X follows a Bernoulli distribution with parameter p , denoted $X \hookrightarrow \mathcal{B}(p)$, if :

$$X(\Omega) = \{0, 1\}$$

and

$$p(X = x) = \begin{cases} p & \text{if } x = 1, \\ 1 - p & \text{if } x = 0. \end{cases}$$

The event $(X = 1)$ represents success, and the event $(X = 0)$ represents failure.

Example 1 : In a coin-toss experiment, tossing a fair coin gives heads with probability $\frac{1}{2}$ and tails with probability $\frac{1}{2}$. If the coin is biased, then heads appears with probability $p \neq \frac{1}{2}$ and tails with probability $q = 1 - p \neq \frac{1}{2}$. By assigning heads the value 1 and tails the value 0, we obtain a Bernoulli distribution.

Proposition 1.2: Expectation and Variance of the Bernoulli Distribution

$$E(X) = p$$

$$V(X) = p(1 - p)$$

1- 2) Binomial Distribution

Parameters : positive integer n and $p \in [0, 1]$.

- 👉 **Experiment :** n independent repetitions of a trial with success probability p .
- 👉 **Sample space :** the set of all sequences of H and T of length n .
- 👉 **Random variable X :** number of successes (Heads) observed in the n trials.
- 👉 **Model of :** number of successes in a fixed number of independent Bernoulli trials.

For example, when $n = 3$, the tree of possible outcomes is :

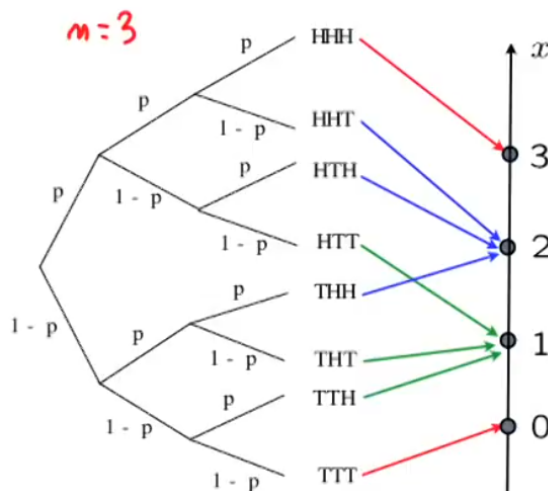


Figure 3.1 – Three independent tosses of a coin with $p(H) = p$.

Since X counts the number of Heads, for $n = 3$ the possible values are 0, 1, 2, 3. For instance,

$$p_X(2) = p(HHT) + p(HTH) + p(THH) = 3p^2(1-p) = \binom{3}{2}p^2(1-p).$$

Probability mass function of a binomial random variable

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \dots, n.$$

Definition 1.3: Binomial Distribution

A random variable X follows a binomial distribution, written

$$X \hookrightarrow \mathcal{B}(n, p),$$

if

- 👉 an experiment is repeated independently n times, each trial resulting in success with probability p ,
- 👉 X counts the total number of successes among the n trials,

$$X(\Omega) = \mathbb{N}.$$

Example 2 : We toss a coin n times. For each i with $1 \leq i \leq n$, define the Bernoulli variable

$$X_i = \begin{cases} 1 & \text{if the } i^{\text{th}} \text{ toss gives Heads,} \\ 0 & \text{otherwise.} \end{cases}$$

Let

$$Y = X_1 + X_2 + \cdots + X_n.$$

Then Y counts the number of Heads observed in the n tosses, and hence

$$Y \hookrightarrow \mathcal{B}(n, p).$$

Theorem 1.4: Relationship Between the Binomial and Bernoulli Distributions

If

$$Y = X_1 + X_2 + \cdots + X_n,$$

where X_1, \dots, X_n are independent Bernoulli random variables with parameter p , then

$$Y \hookrightarrow \mathcal{B}(n, p).$$

To determine the probabilities of the elementary events of a random variable following a binomial law, it is necessary to calculate the number of ways to obtain k successes in n trials.

This means finding the number of (unordered) combinations of k objects among n , with $k \leq n$. These combinations are unordered, since only the total number of successes matters, not their order of occurrence. It is known that the number of ways to obtain k successes and $n - k$ failures is given by the binomial coefficient C_n^k . Multiplying this coefficient by the corresponding probabilities of success and failure yields the binomial law.

We therefore have :

Proposition 1.5

The elementary probabilities of a random variable X following a binomial distribution $\mathcal{B}(n, p)$ are given, for every number of successes $k = 0, 1, \dots, n$, by :

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

Proposition 1.6: Expectation and Variance of the Binomial Distribution

$$E(X) = np \quad V(X) = np(1 - p)$$

Good exercise : prove this proposition !

Property 1.7: Sum

Let X_1 and X_2 be two *independent* random variables following binomial distributions $\mathcal{B}(n_1, p)$ and $\mathcal{B}(n_2, p)$ respectively. Then the sum $X_1 + X_2$ follows a binomial distribution $\mathcal{B}(n_1 + n_2, p)$.

This property is easy to interpret : if X_1 represents the number of successes during n_1 identical and independent trials, and X_2 the number of successes during n_2 independent trials (also independent of the first n_1), with the same probability of success, then $X_1 + X_2$ represents the total number of successes in $n_1 + n_2$ identical and independent trials.

1- 3) Geometric Distribution

Parameter : probability of success p , with $0 < p \leq 1$.

👉 **Experiment** : infinitely many independent tosses of a coin with $p(\text{Heads}) = p$.

👉 **Sample space** : the set of infinite sequences of H and T .

👉 **Random variable X** : number of trials until the first success (first Heads).

👉 **Model of** : waiting times ; number of trials required to obtain the first success.

For example, if the sequence begins as

TTTTHT...

then the first Heads occurs on the 5th toss, so

$$X = 5.$$

Since the first success at trial k requires $(k - 1)$ failures followed by 1 success,

$$p(X = k) = (1 - p)^{k-1}p, \quad k = 1, 2, 3, \dots$$

Probability mass function of a geometric random variable

$$p_X(k) = (1 - p)^{k-1}p, \quad k = 1, 2, 3, \dots$$

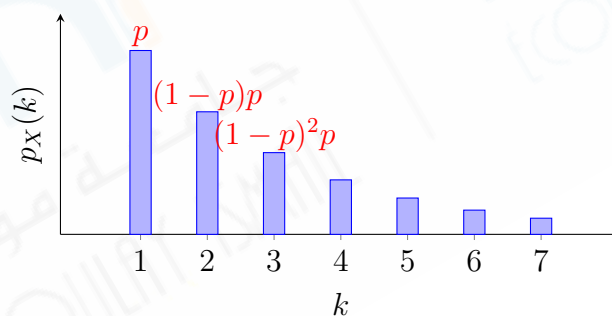


Figure 3.2 – Probability mass function of a geometric distribution.

Definition 1.8: Geometric Distribution

A random variable X follows a geometric distribution, written

$$X \hookrightarrow \mathcal{G}(p),$$

if independent Bernoulli(p) trials are performed and X counts the number of trials until the first success :

$$X(\Omega) = \mathbb{N}^*$$

This is indeed a probability distribution, since :

- ✓ $p(X = k)$ is nonnegative,
- ✓ $\sum_{k \geq 1} p(X = k) = \sum_{k \geq 1} (1-p)^{k-1}p = \frac{p}{1-p} = 1$, where $q = 1 - p \in (0, 1)$.

Proposition 1.9: Expectation and Variance of the Geometric Distribution

$$E(X) = \frac{1}{p}$$

$$V(X) = \frac{q}{p^2} = \frac{1-p}{p^2}$$

Indeed :

$$\checkmark E(X) = \sum_{k \geq 1} k q^{k-1} p = p \sum_{k \geq 0} k q^{k-1} = \frac{p}{(1-q)^2} = \frac{1}{p}.$$

\checkmark To compute $V(X)$, first compute $E(X(X-1))$:

$$\begin{aligned} E(X(X-1)) &= \sum_{k \geq 1} k(k-1) q^{k-1} p \\ &= 2 \frac{pq}{(1-q)^3} \\ &= 2 \frac{q}{p^2}. \end{aligned}$$

Moreover,


$$\begin{aligned} V(X) &= E(X^2) - (E(X))^2 \\ &= E(X(X-1)) + E(X) - (E(X))^2 \\ &= 2 \frac{q}{p^2} + \frac{1}{p} - \frac{1}{p^2}, \end{aligned}$$

hence

$$V(X) = \frac{q}{p^2} = \frac{1-p}{p^2}.$$

Typical Example

Sampling with replacement from an urn containing two types of balls. We observe the index of the first appearance of a given type.

 **Example 3 :** We sample with replacement from an urn containing 113 white balls and 7 black balls. The probability of drawing a black ball is

$$p = \frac{7}{120}.$$

Thus

$$E(X) = \frac{1}{p} = \boxed{\frac{120}{7}},$$

so on average one should expect between 17 and 18 draws to obtain a black ball for the first time.