Notes of Probability and Statistics

Roberto Monte

January 11, 2023

Abstract

These notes are still a work in progress and are intended for internal use. Please, do not cite or quote.

Contents

Part I

Elements of Calculus

Chapter 1

Real Euclidean Spaces

In the following, we will write $\mathbb{N}$ to denote the set of all natural numbers, in the roster form $\mathbb{N} \equiv \{1, 2, 3, \dots\}$, and we will write $\mathbb{N}_0$ for $\mathbb{N} \cup \{0\}$. We will also use the symbols $\mathbb{O}$ and $\mathbb{E}$ to denote the subsets of all odd and even numbers of $\mathbb{N}$, in the set builder form,

$$\mathbb{O} \equiv \{n \in \mathbb{N} : n = 2k - 1,\ k \in \mathbb{N}\} \quad \text{and} \quad \mathbb{E} \equiv \{n \in \mathbb{N} : n = 2k,\ k \in \mathbb{N}\},$$

Referring to $\mathbb{N}_0$ we will also write

$$\mathbb{E}_0 \equiv \{n \in \mathbb{N}_0 : n = 2k,\ k \in \mathbb{N}_0\} = \mathbb{E} \cup \{0\}.$$

We will use the symbol $\mathbb{Z}$ to denote the set of all integer numbers, in the roster form,

$$\mathbb{Z} \equiv \{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\},$$

and the symbol $\mathbb{Q}$ for the set of all rational numbers, in the set builder form,

$$\mathbb{Q} \equiv \left\{x : x = \frac{p}{q},\ p \in \mathbb{Z},\ q \in \mathbb{N},\ \gcd(p, q) = 1\right\},$$

where $\gcd(p, q)$ denotes the *greatest common divisor* of p and q. We will write $\mathbb{R}$ to denote the set of all real numbers. A nice set builder form of $\mathbb{R}$ can be given by using the binary representation of the real numbers in the interval $[0, 1]$, that is,

$$\mathbb{R} \equiv \left\{x : x = z + \xi,\ z \in \mathbb{Z},\ \xi = \sum_{n=1}^{\infty} \frac{\xi_n}{2^n},\ \xi_n = 1 \ \vee\ \xi_n = 0, \quad \forall n \in \mathbb{N}\right\}.$$

we will also use the symbol $\mathbb{R}_+ \equiv \{x \in \mathbb{R} : x \geq 0\}$ [resp. $\mathbb{R}_- \equiv \{x \in \mathbb{R} : x \leq 0\}$, resp. $\mathbb{R}_{++} \equiv \{x \in \mathbb{R} : x > 0\}$, resp. $\mathbb{R}_{--} \equiv \{x \in \mathbb{R} : x < 0\}$] for the set of all *positive* [resp. *negative*, resp. *strictly positive*, resp. *strictly negative*] real numbers. In this context, we recall the absolute value function $|\cdot| : \mathbb{R} \to \mathbb{R}$ given by

$$|x| \stackrel{\text{def}}{=} \begin{cases} x, & \text{if } x \geq 0, \\ -x, & \text{if } x < 0. \end{cases}$$

The absolute value function satisfies the following simple properties.

Proposition 1 *We have*

1. $|x| \in \mathbb{R}_+$, *for every* $x \in \mathbb{R}$ *and* $|x| = 0$, *if and only if* $x = 0$;
2. $|-x| = |x|$, *for every* $x \in \mathbb{R}$;
3. $|xy| = |x|\,|y|$, *for all* $x, y \in \mathbb{R}$.

In addition,

Proposition 2 (Minkowski inequality - convex case) *We have*

$$|x+y|^p \leq 2^{p-1}\left(|x|^p + |y|^p\right), \tag{1.1}$$

for every $p \in [1, +\infty)$ *and for all* $x, y \in \mathbb{R}$. *In particular,*

$$|x+y| \leq |x| + |y|, \tag{1.2}$$

for all $x, y \in \mathbb{R}$.

Proof. The auxiliary function $f : \mathbb{R} \to \mathbb{R}$ given by

$$f(x) \stackrel{\text{def}}{=} |x|^p, \quad \forall x \in \mathbb{R},$$

is convex, for every $p \in [1, \infty)$. Then, we have

$$f\left(\frac{x+y}{2}\right) \leq \frac{1}{2}\left(f(x) + f(y)\right),$$

for all $x, y \in \mathbb{R}$. This implies

$$\left|\frac{x+y}{2}\right|^p \leq \frac{1}{2}\left(|x|^p + |y|^p\right),$$

which is the desired (1.1). □

Proposition 3 (Minkowski inequality - concave case) *For every* $p \in (0, 1)$ *we have*

$$|x+y|^p \leq |x|^p + |y|^p \tag{1.3}$$

for all $x, y \in \mathbb{R}$.

Proof. The auxiliary function $f : \mathbb{R}_+ \to \mathbb{R}$ given by

$$f(x) \stackrel{\text{def}}{=} x^p, \quad \forall x \in \mathbb{R}_+,$$

is concave, for every $p \in (0, 1)$. Then, we have

$$f\left((1-\theta)(x+y)\right) \geq \theta f(0) + (1-\theta) f(x+y)$$

for all $x, y > 0$ and every $\theta \in [0, 1]$. Taking into account that $f(0) = 0$ and choosing

$$\theta \equiv \frac{y}{x+y} \quad \text{and} \quad \theta \equiv \frac{x}{x+y},$$

it follows

$$f(x) \geq \frac{x}{x+y} f(x+y) \quad \text{and} \quad f(y) \geq \frac{y}{x+y} f(x+y).$$

Summing both sides of the above inequalities, we obtain

$$f(x) + f(y) \geq f(x+y),$$

for all $x, y > 0$. That is,

$$(x+y)^p \leq x^p + y^p.$$

It clearly follows

$$(|x| + |y|)^p \leq |x|^p + |y|^p,$$

for all $x, y \in \mathbb{R}$. In the end, since f is increasing, on account of Equation (1.2), we obtain

$$|x+y|^p \leq (|x| + |y|)^p \leq |x|^p + |y|^p,$$

for all $x, y \in \mathbb{R}$. □

Other useful function in a statistic context is the *floor* [resp. *ceiling*] function $\lfloor \cdot \rfloor : \mathbb{R} \to \mathbb{Z}$ [resp. $\lceil \cdot \rceil : \mathbb{R} \to \mathbb{Z}$] given by

$$\lfloor x \rfloor \stackrel{\text{def}}{=} \max \{ z \in \mathbb{Z} : z \leq x \} \quad [\text{resp. } \lceil x \rceil \stackrel{\text{def}}{=} \min \{ z \in \mathbb{Z} : x \leq z \}], \quad \forall x \in \mathbb{R}.$$

We will write $\mathbb{C}$ to denote the set of all complex numbers, in set builder form,

$$\mathbb{C} \equiv \{ z : z = x + iy, \quad \forall x, y \in \mathbb{R} \},$$

where i is the *imaginary unit*. In the context of complex numbers, it is useful to recall the *real part* [resp. *imaginary part*] function $\mathrm{Re} : \mathbb{C} \to \mathbb{R}$ [resp. $\mathrm{Im} : \mathbb{C} \to \mathbb{R}$] given by

$$\mathrm{Re}(z) \stackrel{\text{def}}{=} x, \quad \forall z \in \mathbb{C},\ z = x + iy \quad [\text{resp. } \mathrm{Im}(z) \stackrel{\text{def}}{=} y, \quad \forall z \in \mathbb{C},\ z = x + iy],$$

the *conjugate* function $\overline{\cdot} : \mathbb{C} \to \mathbb{C}$ given by

$$\overline{z} \stackrel{\text{def}}{=} x - iy, \quad \forall z \in \mathbb{C},\ z \equiv x + iy,$$

and the *modulus* function $|\cdot| : \mathbb{C} \to \mathbb{R}_+$ given by

$$|z| \stackrel{\text{def}}{=} \left(x^2 + y^2 \right)^{1/2}, \quad \forall z \in \mathbb{C},\ z = x + iy.$$

Definition 4 *Fixed any $N \in \mathbb{N}$, we call the N-dimensional real space, and we denote it by $\mathbb{R}^N$, the linear space of all column N-tuples of real numbers, that is*

$$\mathbb{R}^N \equiv \{ x : x \equiv (x_1, \ldots, x_N)^\intercal : x_K \in \mathbb{R}, \quad \forall K = 1, \ldots, N \},$$

where $\intercal$ denotes the transpose operator.

Thus, we adopt the standard convention to identify a point $x \in \mathbb{R}^N$ with the column vector $(x_1, \ldots, x_N)^\intercal$ of its entries. Clearly,

$$\mathbb{R}^N \equiv \underbrace{\mathbb{R} \times \cdots \times \mathbb{R}}_{N\text{-times}} \equiv \mathsf{X}_{K=1}^N \mathbb{R},$$

where $\times$ denotes the Cartesian product. In this context, we recall the K-th *canonical projection* of $\mathbb{R}^N$ on $\mathbb{R}$, that is, the function $\pi_K : \mathbb{R}^N \to \mathbb{R}$ given by

$$\pi_K(x) \stackrel{\text{def}}{=} x_K, \quad \forall x \in \mathbb{R}^N,\ x \equiv (x_1, \ldots, x_K, \ldots, x_N)^\intercal. \tag{1.4}$$

where x_K is the Kth *entry* of x, for $K = 1, \ldots, N$, and the *maximum* [resp. *minimum*] functional $\vee : \mathbb{R}^N \to \mathbb{R}$ [resp. $\wedge : \mathbb{R}^N \to \mathbb{R}$] given by

$$\vee(x_1, \ldots, x_N) \stackrel{\text{def}}{=} \max\{x_1, \ldots, x_N\} \quad [\text{resp. } \wedge(x_1, \ldots, x_N) \stackrel{\text{def}}{=} \min\{x_1, \ldots, x_N\}], \quad \forall (x_1, \ldots, x_N)^\intercal \in \mathbb{R}^N.$$

In terms of notation, we will also write

$$\vee(x_1, \ldots, x_N) \equiv \bigvee_{n=1}^{N} x_n \quad \text{and} \quad \wedge(x_1, \ldots, x_N) \equiv \bigwedge_{n=1}^{N} x_n.$$

1.0.1 Fundamental Inequalities on $\mathbb{R}^N$

Let $\mathbb{R}^N$ be the N-dimensional real space, for some $N \in \mathbb{N}$, and let $(x_1, \ldots, x_N)^\intercal, (y_1, \ldots, y_N)^\intercal \in \mathbb{R}^N$.

Proposition 5 (Cauchy-Schwarz inequality) *We have*

$$\sum_{K=1}^{N} |x_K y_K| \leq \left(\sum_{K=1}^{N} x_K^2 \right)^{1/2} \left(\sum_{K=1}^{N} y_K^2 \right)^{1/2}. \tag{1.5}$$

Proof. Since (1.5) is clearly true when at least one between $\sum_{K=1}^N x_K^2$ and $\sum_{K=1}^N y_K^2$ is zero, we prove it under the further assumption $\sum_{K=1}^N x_K^2 \neq 0$ and $\sum_{K=1}^N y_K^2 \neq 0$. In this case, by applying the inequality

$$|xy| \leq \frac{1}{2}\left(x^2 + y^2\right),$$

which holds true for all $x, y \in \mathbb{R}$, we can write

$$\begin{aligned} \frac{|x_K|}{\left(\sum_{L=1}^N x_L^2\right)^{1/2}} \frac{|y_K|}{\left(\sum_{L=1}^N y_L^2\right)^{1/2}} &\leq \frac{1}{2}\left(\left(\frac{|x_K|}{\left(\sum_{L=1}^N x_L^2\right)^{1/2}} \right)^2 + \left(\frac{|y_K|}{\left(\sum_{L=1}^N y_L^2\right)^{1/2}} \right)^2 \right) \\ &= \frac{1}{2}\left(\frac{x_K^2}{\sum_{L=1}^N x_L^2} + \frac{y_K^2}{\sum_{L=1}^N y_L^2} \right), \end{aligned}$$

for every $K = 1, \ldots, N$. Therefore, summing for $K = 1, \ldots, N$, we obtain

$$\begin{aligned} \sum_{K=1}^{N} \frac{|x_K|}{\left(\sum_{L=1}^N x_L^2\right)^{1/2}} \frac{|y_K|}{\left(\sum_{L=1}^N y_L^2\right)^{1/2}} &\leq \sum_{K=1}^{N} \frac{1}{2}\left(\frac{x_K^2}{\sum_{L=1}^N x_L^2} + \frac{y_K^2}{\sum_{L=1}^N y_L^2} \right) \\ &= \frac{1}{2}\left(\sum_{K=1}^{N} \frac{x_K^2}{\sum_{L=1}^N x_L^2} + \sum_{K=1}^{N} \frac{y_K^2}{\sum_{L=1}^N y_L^2} \right) \\ &= \frac{1}{2}\left(\frac{\sum_{K=1}^N x_K^2}{\sum_{L=1}^N x_L^2} + \frac{\sum_{K=1}^N y_K^2}{\sum_{\ell=1}^n b_\ell^2} \right) \\ &= 1. \end{aligned} \tag{1.6}$$

From (1.6) the Cauchy-Schwarz inequality (1.5) immediately follows. □

Let $p \in (1, +\infty)$.

Definition 6 *We call conjugate exponent of p the real number*

$$q \overset{def}{=} p/(p-1). \tag{1.7}$$

Remark 7 *We have clearly $q \in (1, +\infty)$ and*

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Lemma 8 (Hölder inequality) *For every $p \in (1, +\infty)$ with conjugate exponent q we have*

$$|xy| \leq \frac{|x|^p}{p} + \frac{|y|^q}{q}, \tag{1.8}$$

for all $x, y \in \mathbb{R}$. Furthermore, the equality holds in (1.8) if and only if $|x| = |y|^{q/p}$.

Proof. Since (8) is clearly true when at least one between x and y is zero, we prove it under the additional assumption $x \neq 0$ and $y \neq 0$. Now, it is not difficult to show that the auxiliary function $f : \mathbb{R}_+ \to \mathbb{R}$ given by

$$f(u) \overset{\text{def}}{=} \frac{u^p}{p} + \frac{1}{q} - u, \quad \forall u \in \mathbb{R}_+, \tag{1.9}$$

has a unique minimum in $u = 1$, where it takes the zero value. In fact, we have

$$f'(u) = u^{p-1} - 1,$$

for every $u \in \mathbb{R}_+$. On the other hand,

$$p - 1 > 0.$$

Hence,

$$f'(u) \geq 0 \Leftrightarrow u^{p-1} - 1 \geq 0 \Leftrightarrow u \geq 1.$$

Thus, the function f has a unique minimum at the point $u = 1$ and we have

$$f(u) \geq f(1) = 0,$$

for every $u \in \mathbb{R}_+$. As a consequence, we obtain

$$u \leq \frac{u^p}{p} + \frac{1}{q}, \tag{1.10}$$

for every $u \in \mathbb{R}_+$. Furthermore, the equality holds in (1.10) if and only if $u = 1$. Then, choosing $u \equiv |x|\,|y|^{-q/p}$, we obtain

$$|x|\,|y|^{-q/p} \leq \frac{|x|^p\,|y|^{-q}}{p} + \frac{1}{q},$$

which implies

$$|x|\,|y|^{-q/p}\,|y|^{1+q/p} \leq \frac{|x|^p\,|y|^{-q}\,|y|^{1+q/p}}{p} + \frac{|y|^{1+q/p}}{q}.$$

The latter is the desired (1.8) by virtue of the equalities

$$1+\frac{q}{p}-q=1+\frac{p}{(p-1)\,p}-\frac{p}{p-1}=1+\frac{1}{p-1}-\frac{p}{p-1}=0$$

and

$$1+q/p=1+\frac{p}{(p-1)\,p}=1+\frac{1}{p-1}=q.$$

In the end, the equality holds if and only if

$$|x|\,|y|^{-q/p}=1$$

as claimed. □

Proposition 9 (Hölder inequality) *For every $p\in(1,+\infty)$ with conjugate exponent q, we have*

$$\sum_{K=1}^{N}|x_k y_k|\leq\left(\sum_{K=1}^{N}|x_K|^p\right)^{1/p}\left(\sum_{K=1}^{N}|y_K|^q\right)^{1/q}. \tag{1.11}$$

Proof. Since the Hölder inequality (1.11) is clearly true when at least one between $\sum_{K=1}^{N}|x_K|^p$ and $\sum_{K=1}^{N}|y_K|^q$ is zero, we prove it under the further assumption $\sum_{K=1}^{N}|x_K|^p\neq 0$ and $\sum_{K=1}^{N}|y_K|^q\neq 0$. In this case, setting $u\equiv\left(\sum_{K=1}^{N}|x_K|^p\right)^{1/p}$ and $v\equiv\left(\sum_{K=1}^{N}|y_K|^q\right)^{1/q}$, thanks to Equation (**??**), we can write

$$\frac{|x_K y_K|}{uv}\leq\frac{|x_K|^p}{pu^P}+\frac{|y_K|^q}{qv^q}$$

for every $K=1,\dots,N$. Therefore, summing for $K=1,\dots,N$, we obtain

$$\begin{aligned}\sum_{K=1}^{N}\frac{|x_K y_K|}{uv}&\leq\sum_{K=1}^{N}\left(\frac{|x_K|^p}{pu^P}+\frac{|y_K|^q}{qv^q}\right)\\&=\frac{\sum_{K=1}^{N}|x_K|^p}{pu^P}+\frac{\sum_{K=1}^{N}|y_K|^q}{qv^q}\\&=\frac{u^P}{pu^P}+\frac{v^q}{qv^q}\\&=1.\end{aligned}$$

It then follows

$$\sum_{K=1}^{N}|x_K y_K|\leq uv=\left(\sum_{K=1}^{N}|x_K|^p\right)^{1/p}\left(\sum_{K=1}^{N}|y_K|^q\right)^{1/q},$$

as desired. □

Note that in case $p=2$, we have $q=2$ and the Hölder inequality (1.11) becomes the Cauchy-Schwarz inequality (1.5).

Proposition 10 (Minkowski inequality) *We have*

$$\left(\sum_{K=1}^{N}|x_K+y_K|^p\right)^{1/p}\leq\left(\sum_{K=1}^{N}|x_K|^p\right)^{1/p}+\left(\sum_{K=1}^{N}|y_K|^p\right)^{1/p}. \tag{1.12}$$

Proof. Since the Minkowski inequality (1.12) is clearly true when $\sum_{K=1}^{N} |x_K + y_K| > 0$, we prove it under the further assumption $\sum_{K=1}^{N} |x_K + y_K| > 0$. In this case, we can clearly write

$$\begin{aligned}\sum_{K=1}^{N} |x_K + y_K|^p &= \sum_{K=1}^{N} |x_K + y_K| \, |x_K + y_K|^{p-1} \\ &\leq \sum_{K=1}^{N} (|x_K| + |y_K|) \, |x_K + y_K|^{p-1} \\ &= \sum_{K=1}^{N} |x_K| \, |x_K + y_K|^{p-1} + \sum_{K=1}^{N} |y_K| \, |x_K + y_K|^{p-1} . \end{aligned} \tag{1.13}$$

On the other hand, thanks to the Hölder inequality (1.11), we have

$$\sum_{K=1}^{N} |x_K| \, |x_K + y_K|^{p-1} \leq \left(\sum_{K=1}^{N} |x_K|^p \right)^{1/p} \left(\sum_{K=1}^{N} |x_K + y_K|^{(p-1)q} \right)^{1/q} = \left(\sum_{K=1}^{N} |x_K|^p \right)^{1/p} \left(\sum_{K=1}^{N} |x_K + y_K|^p \right)^{1/q} . \tag{1.14}$$

Similarly,

$$\sum_{K=1}^{N} |y_K| \, |x_K + y_K|^{p-1} \leq \left(\sum_{K=1}^{N} |y_K|^p \right)^{1/p} \left(\sum_{K=1}^{N} |x_K + y_K|^{(p-1)q} \right)^{1/q} = \left(\sum_{K=1}^{N} |y_K|^p \right)^{1/p} \left(\sum_{K=1}^{N} |x_K + y_K|^p \right)^{1/q} \tag{1.15}$$

Combining (1.13)-(1.15), it follows

$$\sum_{K=1}^{N} |x_K + y_K|^p \leq \left(\sum_{K=1}^{N} |x_K|^p \right)^{1/p} \left(\sum_{K=1}^{N} |x_K + y_K|^p \right)^{1/q} + \left(\sum_{K=1}^{N} |y_K|^p \right)^{1/p} \left(\sum_{K=1}^{N} |x_K + y_K|^p \right)^{1/q} \tag{1.16}$$

Due to the assumption $\sum_{K=1}^{N} |x_K + y_K| > 0$, which implies that there exists at least a $K_0 \in \{1, \ldots, N\}$ such that $|x_{K_0} + y_{K_0}| > 0$, Equation (1.16) implies

$$\frac{\sum_{K=1}^{N} |x_K + y_K|^p}{\left(\sum_{K=1}^{N} |x_K + y_K|^p \right)^{1/q}} \leq \left(\sum_{K=1}^{N} |x_K|^p \right)^{1/p} + \left(\sum_{K=1}^{N} |y_K|^p \right)^{1/p} . \tag{1.17}$$

On the other hand, since

$$1 - 1/q = 1 - \frac{p-1}{p} = \frac{1}{p},$$

we have

$$\frac{\sum_{K=1}^{N} |x_K + y_K|^p}{\left(\sum_{K=1}^{N} |x_K + y_K|^p \right)^{1/q}} = \left(\sum_{K=1}^{N} |x_K + y_K|^p \right)^{1-1/q} = \left(\sum_{K=1}^{N} |x_K + y_K|^p \right)^{1/p} . \tag{1.18}$$

In the end, combining (1.17) and (1.18), the desired (1.12) follows. □

1.0.2 Topology on $\mathbb{R}^N$

Let $\mathbb{R}^N$ be the N-dimensional real space, for some $N \in \mathbb{N}$, (see Definition 4).

Definition 11 *We call the Euclidean distance on $\mathbb{R}^N$ the map $d : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}_+$ given by*

$$d(x,y) \overset{def}{=} \left(\sum_{K=1}^{N} (x_n - y_n)^2 \right)^{1/2}, \quad \forall x, y \in \mathbb{R}^N,$$

where $x \equiv (x_1, \ldots, x_N)^\intercal$ and $y \equiv (y_1, \ldots, y_N)^\intercal$.

Remark 12 *If $N = 1$*

$$d(x,y) = |x - y|,$$

for all $x, y \in \mathbb{R}$.

Proposition 13 *We have*

1. *$d(x,y) = d(y,x)$, for all $x, y \in \mathbb{R}^N$;*
2. *$d(x,y) = 0 \Leftrightarrow x = y$;*
3. *$d(x,z) \leq d(x,y) + d(y,z)$, for all $x, y \in \mathbb{R}^N$.*

Proof. . □

Definition 14 *We call the* Euclidean norm *on $\mathbb{R}^N$ the map $\|\cdot\|_2 : \mathbb{R}^N \to \mathbb{R}_+$ given by*

$$\|x\|_2 \overset{def}{=} \left(\sum_{n=1}^{N} x_n^2 \right)^{1/2}, \quad \forall x \in \mathbb{R}^N,$$

where $x \equiv (x_1, \ldots, x_N)^\intercal$.

Remark 15 *If $N = 1$*

$$\|x\|_2 = |x|,$$

for every $x \in \mathbb{R}$.

Remark 16 *We have*

$$d(x,y) = \|x - y\|_2,$$

for all $x, y \in \mathbb{R}^N$

Proposition 17 *We have*

1. *$\|x\|_2 \geq 0$, for every $x \in \mathbb{R}^N$ and $\|x\|_2 = 0$, if and only if $x = 0$;*
2. *$\|\alpha x\|_2 = |\alpha| \, \|x\|_2$, for every $\alpha \in \mathbb{R}$ and $x \in \mathbb{R}^N$;*
3. *$\|x - y\|_2 \leq \|x - z\|_2 + \|z - y\|_2$, for all $x, y, z \in \mathbb{R}^N$;*
4. *$|\|x\|_2 - \|y\|_2| \leq \|x - y\|_2$, for all $x, y \in \mathbb{R}^N$.*

Definition 18 *We call the N-dimensional real space $\mathbb{R}^N$* Euclidean*, when $\mathbb{R}^N$ is endowed with the Euclidean distance and norm.*

Definition 19 *We call the* sum norm *on $\mathbb{R}^N$ the map $\|\cdot\|_2 : \mathbb{R}^N \to \mathbb{R}_+$ given by*

$$\|x\|_{2,\ \mathrm{sum}} \stackrel{def}{=} \sum_{n=1}^{N} |x_n| , \quad \forall x \in \mathbb{R}^N.$$

Proposition 20 *The sum norm on $\mathbb{R}^N$ satisfies 1.-4. in Proposition 17. In addition, we have*

$$\|x\|_{2,\ \mathrm{sum}} \leq N \|\cdot\|_2 \quad \textit{and} \quad \|\cdot\|_2 \leq \|x\|_{2,\ \mathrm{sum}} , \tag{1.19}$$

for every $x \in \mathbb{R}^N$.

Proof. To prove Equation (1.19), observe that we have

$$x_n^2 \leq \sum_{n=1}^{N} x_n^2,$$

for every $x \in \mathbb{R}^N$ and $n = 1, \ldots, N$. It follows

$$|x_n| = \left(x_n^2\right)^{1/2} \leq \left(\sum_{n=1}^{N} x_n^2 \right)^{1/2} ,$$

for every $x \in \mathbb{R}^N$ and $n = 1, \ldots, N$. This implies

$$\|x\|_{2,\ \mathrm{sum}} = \sum_{n=1}^{N} |x_n| \leq \sum_{n=1}^{N} \left(\sum_{n=1}^{N} x_n^2 \right)^{1/2} = N \left(\sum_{n=1}^{N} x_n^2 \right)^{1/2} = N \|x\|_2 ,$$

for every $x \in \mathbb{R}^N$. On the other hand, we have

$$\sum_{n=1}^{N} x_n^2 \leq \left(\sum_{n=1}^{N} |x_n| \right)^2 ,$$

for every $x \in \mathbb{R}^N$. This implies

$$\|x\|_2 = \left(\sum_{n=1}^{N} x_n^2 \right)^{1/2} \leq \sum_{n=1}^{N} |x_n| = \|x\|_{2,\ \mathrm{sum}} ,$$

for every $x \in \mathbb{R}^N$, and completes the proof. □

Definition 21 *We call the* maximum norm *on $\mathbb{R}^N$ the map $\|\cdot\|_{2,\max} : \mathbb{R}^N \to \mathbb{R}_+$ given by*

$$\|x\|_{2,\ \max} \stackrel{def}{=} \max_{n=1,\ldots,N} \{|x_n|\} , \quad \forall x \in \mathbb{R}^N.$$

Proposition 22 *The maximum norm on $\mathbb{R}^N$ satisfies 1.-4. in Proposition 17. In addition, we have*

$$\|x\|_{2,\ \max} \leq \|\cdot\|_2 \quad \text{and} \quad \|\cdot\|_2 \leq N \|x\|_{2,\ \max}, \tag{1.20}$$

for every $x \in \mathbb{R}^N$.

Proof. To prove Equation (1.20), observe that we have

$$|x_n| = \left(x_n^2\right)^{1/2} \leq \left(\sum_{n=1}^{N} x_n^2\right)^{1/2},$$

for every $x \in \mathbb{R}^N$ and $n = 1, \ldots, N$. This implies

$$\|x\|_{2,\ \max} = \max_{n=1,\ldots,N} \{|x_n|\} \leq \left(\sum_{n=1}^{N} x_n^2\right)^{1/2},$$

for every $x \in \mathbb{R}^N$. On the other hand, considering Equation (1.19), we can write

$$\|x\|_{2,\ \max} \leq \left(\sum_{n=1}^{N} x_n^2\right)^{1/2} \leq \sum_{n=1}^{N} |x_n| \leq \sum_{n=1}^{N} \max_{m=1,\ldots,N} \{|x_m|\} = N \max_{n=1,\ldots,N} \{|x_n|\} = N \|x\|_{2,\ \max},$$

for every $x \in \mathbb{R}^N$, which completes the proof.

From Propositions 20 and 22 it follows that the sum and maximum norm on $\mathbb{R}^N$ are equivalent norm on $\mathbb{R}^N$ in the sense that they can be used to define distances on $\mathbb{R}^N$ which are equivalent to the Eucliden distance.

Definition 23 *Given any point $x \in \mathbb{R}^N$ and any $r \geq 0$, we call the subset $B^N(x;r)$ [resp. $D^N(x;r)$] of $\mathbb{R}^N$ given by*

$$B^N(x;r) \overset{def}{=} \left\{y \in \mathbb{R}^N : d(y,x) < r\right\} \qquad [\text{resp. } D^N(x;r) \overset{def}{=} \left\{y \in \mathbb{R}^N : d(y,x) \leq r\right\}],$$

the open *[resp.* closed*]* disk *centered at x with radius r.*

Remark 24 *If $r = 0$ we have*

$$B^N(x;0) = \varnothing \qquad \text{and} \qquad D^N(x;0) = \{x\},$$

for every $x \in \mathbb{R}^N$.

Remark 25 *If $N = 1$, setting $B^1(x;r) \equiv B(x;r)$ and $D^1(x;r) \equiv B(x;r)$, we have*

$$B(x;r) = (x - r, x + r) \qquad \text{and} \qquad D(x;r) = [x - r, x + r],$$

for every $x \in \mathbb{R}$ and every $r \geq 0$.

Definition 26 *We say that a subset O of $\mathbb{R}^N$ is* open *(with respect to the Euclidean distance) if for every $x \in O$ there exists $r > 0$ such that $B^N(x;r) \subseteq O$.*

Example 27 *The subsets $\varnothing$ and $\mathbb{R}^N$ of $\mathbb{R}^N$ are open.*

Example 28 *For any $x \in \mathbb{R}^N$ the singleton $\{x\}$ is not open.*

Discussion. . □

Example 29 *Given any point $x \in \mathbb{R}^N$ and any $r > 0$ the open disk centered at x with radius r is open.*

Discussion. . □

Definition 30 *We call the Euclidean topology of $\mathbb{R}^N$, and we denote it by $\mathcal{O}\left(\mathbb{R}^N\right)$, the family of all subsets of $\mathbb{R}^N$ which are open (with respect to the Euclidean distance). In symbols*

$$\mathcal{O}\left(\mathbb{R}^N\right) \stackrel{def}{=} \left\{O \in \mathcal{P}\left(\mathbb{R}^N\right) : \forall x \in O \ \exists r > 0 : B^N\left(x; r\right) \subseteq O\right\}.$$

where $\mathcal{P}\left(\mathbb{R}^N\right)$ is the power set *of $\mathbb{R}^N$, that is, the family of all subsets of $\mathbb{R}^N$.*

Proposition 31 *The following properties are fulfilled.*

1. *$\varnothing, \mathbb{R}^N \in \mathcal{O}\left(\mathbb{R}^N\right)$;*
2. *$\bigcup_{j\in J} O_j \in \mathcal{O}\left(\mathbb{R}^N\right)$, for any family $\{O_j\}_{j\in J}$ of elements of $\mathcal{O}\left(\mathbb{R}^N\right)$, where J is any set of indices;*
3. *$\bigcap_{k=1}^{n} O_k \in \mathcal{O}\left(\mathbb{R}^N\right)$, for any finite family $\{O_k\}_{k=1}^{n}$ of elements of $\mathcal{O}\left(\mathbb{R}^N\right)$, where n is any natural number.*

Proof. . □

Definition 32 *We say that a subset C of $\mathbb{R}^N$ is* closed *(with respect to the Euclidean distance) if $C^c_{\mathbb{R}^N}$ is open.*

Example 33 *The subsets $\varnothing$ and $\mathbb{R}^N$ of $\mathbb{R}^N$ are closed.*

Example 34 *For any $x \in \mathbb{R}^N$ the singleton $\{x\}$ is closed.*

Discussion. . □

Example 35 *Given any point $x \in \mathbb{R}^N$ and any $r > 0$ the closed disk centered at x with radius r is closed.*

Discussion. . □

Let $\mathcal{C}\left(\mathbb{R}^N\right)$ be the family of all subsets of $\mathbb{R}^N$ which are closed according to the Euclidean distance.

Proposition 36 *The following properties are fulfilled*

1. *$\varnothing, \mathbb{R}^N \in \mathcal{C}\left(\mathbb{R}^N\right)$;*
2. *$\bigcap_{j\in J} C_j \in \mathcal{C}\left(\mathbb{R}^N\right)$, for any family $\{C_j\}_{j\in J}$ of elements of $\mathcal{C}\left(\mathbb{R}^N\right)$, where J is any set of indices;*
3. *$\bigcup_{k=1}^{n} C_k \in \mathcal{C}\left(\mathbb{R}^N\right)$, for any finite family $\{C_k\}_{k=1}^{n}$ of elements of $\mathcal{C}\left(\mathbb{R}^N\right)$, where n is any natural number.*

Proposition 37 *The sets $\varnothing$ and $\mathbb{R}^N$ are the only subsets of $\mathbb{R}^N$ which are both open and closed.*

Proof. . □

Example 38 *For any $a, b \in \mathbb{R}$ such that $a < b$, the intervals $(a, b]$ and $[a, b)$ of $\mathbb{R}$ are neither open nor closed.*

Discussion. . □

1.0.3 Borel σ-algebra of $\mathbb{R}^N$

Let $\mathbb{R}^N$ be the N-dimensional real Euclidean space, for some $N \in \mathbb{N}$, (see Definition 18), and let $\mathcal{P}\left(\mathbb{R}^N\right)$ be the power set of $\mathbb{R}^N$.

Definition 39 *We call a* σ-algebra *of* $\mathbb{R}^N$ *any non-empty subfamily* $\mathcal{S}$ *of* $\mathcal{P}\left(\mathbb{R}^N\right)$ *which satisfies the following properties*

1. $S^c \in \mathcal{S}$, *for every* $S \in \mathcal{S}$;
2. $\bigcup_{n=1}^{\infty} S_n \in \mathcal{S}$. *for every sequence* $(S_n)_{n=1}^{\infty}$ *in* $\mathcal{S}$.

Remark 40 *The power set* $\mathcal{P}\left(\mathbb{R}^N\right)$ *is itself a* σ*-algebra of* $\mathbb{R}^N$ *and the subfamily* $\{\varnothing, \Omega\}$ *of* $\mathcal{P}\left(\mathbb{R}^N\right)$ *is also a* σ*-algebra of* $\mathbb{R}^N$.

Definition 41 *The* σ*-algebra* $\mathcal{P}\left(\mathbb{R}^N\right)$ *[resp.* $\{\varnothing, \Omega\}$*] is called the* discrete *[resp.* trivial*]* σ*-algebra of* $\mathbb{R}^N$.

Remark 42 *With respect to the inclusion relationship on the class of all families of subsets of* $\mathbb{R}^N$, *the discrete* σ*-algebra* $\mathcal{P}\left(\mathbb{R}^N\right)$ *characterizes as the largest* σ*-algebra of* $\mathbb{R}^N$ *and the trivial* σ*-algebra* $\{\varnothing, \Omega\}$ *characterizes as the smallest* σ*-algebra of* $\mathbb{R}^N$. *In symbols,*

$$\{\varnothing, \Omega\} \subseteq \mathcal{S} \subseteq \mathcal{P}\left(\mathbb{R}^N\right),$$

for every σ*-algebra* $\mathcal{S}$ *on* $\mathbb{R}^N$.

Example 43 *The family* $\mathcal{S}_{\text{count-cocount}}$ *of all subsets of* $\mathbb{R}^N$ *which are countable or have countable complement, in symbols*

$$\mathcal{S}_{\text{count-cocount}} \equiv \{S \in \mathcal{P}\left(\mathbb{R}^N\right) : |S| \leq \aleph_0 \vee |S^c| \leq \aleph_0\},$$

is a σ*-algebra such that*

$$\{\varnothing, \Omega\} \subset \mathcal{S} \subset \mathcal{P}\left(\mathbb{R}^N\right).$$

Proposition 44 *Given any non-empty subfamily* $\mathcal{B}$ *of* $\mathcal{P}\left(\mathbb{R}^N\right)$ *there exists a* σ*-algebra of* $\mathbb{R}^N$ *which characterizes as the smallest* σ*-algebra of* $\mathbb{R}^N$ *containing* $\mathcal{B}$. *Clearly, such a* σ*-algebra is unique.*

Proof. Let $\mathfrak{B}$ the class of all σ-algebras on $\mathbb{R}^N$ containing $\mathcal{B}$. Then $\mathfrak{B}$ is non-empty, since $\mathcal{P}\left(\mathbb{R}^N\right) \in \mathfrak{B}$. Hence, consider the family $\bigcap_{\mathcal{S}\in\mathfrak{B}} \mathcal{S}$. Such a family is a σ-algebra of $\mathbb{R}^N$, since it clearly satisfies Properties 1 and 2. In addition, $\mathcal{B} \subseteq \bigcap_{\mathcal{S}\in\mathfrak{B}} \mathcal{S}$, since $\mathcal{B} \subseteq \mathcal{S}$, for every $\mathcal{S} \in \mathfrak{B}$. As a consequence $\bigcap_{\mathcal{S}\in\mathfrak{B}} \mathcal{S}$ itself is in $\mathfrak{B}$. This implies that $\bigcap_{\mathcal{S}\in\mathfrak{B}} \mathcal{S}$ is the smallest σ-algebra of $\mathbb{R}^N$ containing $\mathcal{B}$. □

Definition 45 *We call the* σ*-algebra presented in Proposition 44 the* σ*-algebra generated by* $\mathcal{B}$ *and we denote it by the symbol* $\sigma(\mathcal{B})$.

Definition 46 *We say that a non-empty subfamily* $\mathcal{B}$ *of* $\mathcal{P}\left(\mathbb{R}^N\right)$ generates *a* σ*-algebra* $\mathcal{S}$ *on* $\mathbb{R}^N$ *or* $\mathcal{B}$ *is a* basis *for* $\mathcal{S}$, *if*

$$\sigma(\mathcal{B}) = \mathcal{S}.$$

Different non-empty subfamilies of $\mathcal{P}\left(\mathbb{R}^N\right)$ can generate the same σ-algebra of $\mathbb{R}^N$.

Example 47 *The families of all open and closed subsets of $\mathbb{R}^N$ (with respect to the Euclidean distance) generate the same σ-algebra of $\mathbb{R}^N$. In symbols,*

$$\sigma\left(\mathcal{O}\left(\mathbb{R}^N\right)\right) = \sigma\left(\mathcal{C}\left(\mathbb{R}^N\right)\right).$$

Definition 48 *We call the σ-algebra of $\mathbb{R}^N$ generated by the family of all open (closed) subsets of $\mathbb{R}^N$ the Borel σ-algebra of $\mathbb{R}^N$ and we denote it by the symbol $\mathcal{B}\left(\mathbb{R}^N\right)$.*

Proposition 49 *Each of the following families of subsets of $\mathbb{R}^N$ is a basis for $\mathcal{B}(\mathbb{R}^N)$*

1. *the family $\mathcal{I}_{o.\mathbb{R}}\left(\mathbb{R}^N\right)$ of all open intervals of $\mathbb{R}^N$ with real endpoints*

$$\mathcal{I}_{o.\mathbb{R}}\left(\mathbb{R}^N\right) \equiv \{I \in \mathcal{P}\left(\mathbb{R}^N\right) : I = \mathsf{X}_{K=1}^N(a_K, b_K), \quad a_K, b_K \in \mathbb{R},\ a_K \leq b_K, \quad \forall K = 1, \ldots, N\};$$

2. *the family $\mathcal{I}_{c.\mathbb{R}}\left(\mathbb{R}^N\right)$ of all closed intervals of $\mathbb{R}^N$ with real endpoints*

$$\mathcal{I}_{c.\mathbb{R}}\left(\mathbb{R}^N\right) \equiv \{I \in \mathcal{P}\left(\mathbb{R}^N\right) : I = \mathsf{X}_{K=1}^N [a_K, b_K], \quad a_K, b_K \in \mathbb{R},\ a_K \leq b_K, \quad \forall K = 1, \ldots, N\};$$

3. *the family $\mathcal{I}_{o.c.\mathbb{R}}\left(\mathbb{R}^N\right)$ of all open-closed intervals of $\mathbb{R}^N$ with real endpoints*

$$\mathcal{I}_{o.c.\mathbb{R}}\left(\mathbb{R}^N\right) \equiv \{I \in \mathcal{P}\left(\mathbb{R}^N\right) : I = \mathsf{X}_{K=1}^N(a_K, b_K], \quad a_K, b_K \in \mathbb{R},\ a_K \leq b_K, \quad \forall K = 1, \ldots, N\};$$

4. *the family $\mathcal{I}_{c.o.\mathbb{R}}\left(\mathbb{R}^N\right)$ of all closed-open intervals of $\mathbb{R}^N$ with real endpoints*

$$\mathcal{I}_{c.o.\mathbb{R}}\left(\mathbb{R}^N\right) \equiv \{I \in \mathcal{P}\left(\mathbb{R}^N\right) : I = \mathsf{X}_{K=1}^N[a_K, b_K), \quad a_K, b_K \in \mathbb{R},\ a_K \leq b_K, \quad \forall K = 1, \ldots, N\}.$$

Proof. . □

Proposition 50 *Each of the following families of subsets of $\mathbb{R}^N$ is a basis for $\mathcal{B}(\mathbb{R}^N)$*

1. *the family $\mathcal{I}_{o.\mathbb{Q}}\left(\mathbb{R}^N\right)$ of all open intervals of $\mathbb{R}^N$ with rational endpoints*

$$\mathcal{I}_{o.\mathbb{Q}}\left(\mathbb{R}^N\right) \equiv \{I \in \mathcal{P}\left(\mathbb{R}^N\right) : I = \mathsf{X}_{K=1}^N(a_K, b_K), \quad a_K, b_K \in \mathbb{Q},\ a_K \leq b_K, \quad \forall K = 1, \ldots, N\};$$

2. *the family $\mathcal{I}_{c.\mathbb{Q}}\left(\mathbb{R}^N\right)$ of all closed intervals of $\mathbb{R}^N$ with rational endpoints*

$$\mathcal{I}_{c.\mathbb{Q}}\left(\mathbb{R}^N\right) \equiv \{I \in \mathcal{P}\left(\mathbb{R}^N\right) : I = \mathsf{X}_{K=1}^N [a_K, b_K], \quad a_K, b_K \in \mathbb{Q},\ a_K \leq b_K, \quad \forall K = 1, \ldots, N\};$$

3. *the family $\mathcal{I}_{o.c.\mathbb{Q}}\left(\mathbb{R}^N\right)$ of all open-closed intervals of $\mathbb{R}^N$ with rational endpoints*

$$\mathcal{I}_{o.c.\mathbb{Q}}\left(\mathbb{R}^N\right) \equiv \{I \in \mathcal{P}\left(\mathbb{R}^N\right) : I = \mathsf{X}_{K=1}^N(a_K, b_K], \quad a_K, b_K \in \mathbb{Q},\ a_K \leq b_K, \quad \forall K = 1, \ldots, N\};$$

4. *the family $\mathcal{I}_{c.o.\mathbb{Q}}\left(\mathbb{R}^N\right)$ of all closed-open intervals of $\mathbb{R}^N$ with rational endpoints*

$$\mathcal{I}_{c.o.\mathbb{Q}}\left(\mathbb{R}^N\right) \equiv \{I \in \mathcal{P}\left(\mathbb{R}^N\right) : I = \mathsf{X}_{K=1}^N[a_K, b_K), \quad a_K, b_K \in \mathbb{Q},\ a_K \leq b_K, \quad \forall K = 1, \ldots, N\}.$$

Proof. . □

Proposition 51 *Each of the following families of subsets of $\mathbb{R}^N$ is a basis for $\mathcal{B}(\mathbb{R}^N)$*

1. *the family $\mathcal{H}_{r.o.\mathbb{R}}\left(\mathbb{R}^N\right)$ of all right-open half-lines of $\mathbb{R}^N$ with rational endpoint*
$$\mathcal{H}_{r.o.\mathbb{R}}\left(\mathbb{R}^N\right) = \{H \in \mathcal{P}\left(\mathbb{R}^N\right) : H = \mathsf{X}_{K=1}^{N}(-\infty, a_K), \quad a_K \in \mathbb{R}, \quad \forall K = 1, \ldots, N\};$$
2. *the family $\mathcal{H}_{r.c.\mathbb{R}}\left(\mathbb{R}^N\right)$ of all right-closed half-lines of $\mathbb{R}^K$ with rational endpoint*
$$\mathcal{H}_{r.c.\mathbb{R}}\left(\mathbb{R}^N\right) = \{H \in \mathcal{P}\left(\mathbb{R}^N\right) : H = \mathsf{X}_{K=1}^{N}(-\infty, a_K], \quad a_K \in \mathbb{R}, \quad \forall K = 1, \ldots, N\};$$
3. *the family $\mathcal{H}_{r.c.\mathbb{R}}\left(\mathbb{R}^N\right)$ of all left-open half-lines of $\mathbb{R}^N$ with rational endpoint*
$$\mathcal{H}_{l.o.\mathbb{R}}\left(\mathbb{R}^N\right) = \{H \in \mathcal{P}\left(\mathbb{R}^N\right) : H = \mathsf{X}_{K=1}^{N}(a_K, -\infty), \quad a_K \in \mathbb{R}, \quad \forall K = 1, \ldots, N\};$$
4. *the family $\mathcal{H}_{r.c.\mathbb{R}}\left(\mathbb{R}^N\right)$ of all left-closed half-lines of $\mathbb{R}^N$ with rational endpoint*
$$\mathcal{H}_{l.o.\mathbb{R}}\left(\mathbb{R}^N\right) = \{H \in \mathcal{P}\left(\mathbb{R}^N\right) : H = \mathsf{X}_{K=1}^{N}[a_K, -\infty), \quad a_K \in \mathbb{R}, \quad \forall K = 1, \ldots, N\}.$$

Proof. . □

Proposition 52 *Each of the following families of subsets of $\mathbb{R}^N$ is a basis for $\mathcal{B}(\mathbb{R}^N)$*

1. *the family $\mathcal{H}_{r.o.\mathbb{Q}}\left(\mathbb{R}^N\right)$ of all right-open half-lines of $\mathbb{R}^N$ with rational endpoint*
$$\mathcal{H}_{r.o.\mathbb{Q}}\left(\mathbb{R}^N\right) = \{H \in \mathcal{P}\left(\mathbb{R}^N\right) : H = \mathsf{X}_{K=1}^{N}(-\infty, a_K), \quad a_K \in \mathbb{Q}, \quad \forall K = 1, \ldots, N\};$$
2. *the family $\mathcal{H}_{r.c.\mathbb{Q}}\left(\mathbb{R}^N\right)$ of all right-closed half-lines of $\mathbb{R}^K$ with rational endpoint*
$$\mathcal{H}_{r.c.\mathbb{Q}}\left(\mathbb{R}^N\right) = \{H \in \mathcal{P}\left(\mathbb{R}^N\right) : H = \mathsf{X}_{K=1}^{N}(-\infty, a_K], \quad a_K \in \mathbb{Q}, \quad \forall K = 1, \ldots, N\};$$
3. *the family $\mathcal{H}_{r.c.\mathbb{Q}}\left(\mathbb{R}^N\right)$ of all left-open half-lines of $\mathbb{R}^N$ with rational endpoint*
$$\mathcal{H}_{l.o.\mathbb{Q}}\left(\mathbb{R}^N\right) = \{H \in \mathcal{P}\left(\mathbb{R}^N\right) : H = \mathsf{X}_{K=1}^{N}(a_K, -\infty), \quad a_K \in \mathbb{Q}, \quad \forall K = 1, \ldots, N\};$$
4. *the family $\mathcal{H}_{r.c.\mathbb{Q}}\left(\mathbb{R}^N\right)$ of all left-closed half-lines of $\mathbb{R}^N$ with rational endpoint*
$$\mathcal{H}_{l.o.\mathbb{Q}}\left(\mathbb{R}^N\right) = \{H \in \mathcal{P}\left(\mathbb{R}^N\right) : H = \mathsf{X}_{K=1}^{N}[a_K, -\infty), \quad a_K \in \mathbb{Q}, \quad \forall K = 1, \ldots, N\}.$$

Proof. . □

Proposition 53 *Assume $\mathcal{B}$ is a basis for $\mathcal{B}(\mathbb{R})$ and consider the family*
$$\mathcal{B}^N \equiv \{B \in \mathcal{P}\left(\mathbb{R}^N\right) : B = \mathsf{X}_{K=1}^{N} B_K, \quad B_K \in \mathcal{B}, \quad \forall K = 1, \ldots, N\}.$$
Then, $\mathcal{B}^N$ is a basis for $\mathcal{B}(\mathbb{R}^N)$.

Proof. . □

Proposition 54 *The subset of $\mathbb{R}^N$*
$$x + B \equiv \left\{y \in \mathbb{R}^N : y = x + b,\ b \in B\right\}$$
is in $\mathcal{B}\left(\mathbb{R}^N\right)$, for every $x \in \mathbb{R}^N$ and every $B \in \mathcal{B}\left(\mathbb{R}^N\right)$.

Proof. . □

Theorem 55 *The Borel σ-algebra of $\mathbb{R}^N$ is strictly contained in the discrete σ-algebra. In symbols,*
$$\mathcal{B}(\mathbb{R}^N) \subset \mathcal{P}(\mathbb{R}^N).$$

Proof. . □

1.0.4 Borel-Lebesgue Measure on $\mathcal{B}(\mathbb{R}^N)$

We will write $\bar{\mathbb{R}}$ [resp. $\bar{\mathbb{R}}_+$] for the set of all positive real numbers extented with the addiction of the symbols $-\infty$ and $+\infty$ [resp. the symbol $+\infty$], that is $\bar{\mathbb{R}} \equiv \mathbb{R} \cup \{-\infty, +\infty\}$ [resp. $\bar{\mathbb{R}}_+ \equiv \mathbb{R}_+ \cup \{+\infty\}$].

Let $\mathbb{R}^N$ be the N-dimensional real Euclidean space, for some $N \in \mathbb{N}$, and let $\mathcal{B}\left(\mathbb{R}^N\right)$ be the Borel σ-algebra of $\mathbb{R}^N$.

Theorem 56 *There exists a unique function $\mu_L^N : \mathcal{B}\left(\mathbb{R}^N\right) \to \mathbb{R}_+$ with the following properties*

1. *$\mu_L^N\left(\bigcup_{n=1}^{\infty} B_n\right) = \sum_{n=1}^{\infty} \mu_L^N(B_n)$, for every sequence $(B_n)_{n=1}^{\infty}$ of pairwise disjoint sets in $\mathcal{B}\left(\mathbb{R}^N\right)$;*
2. *$\mu_L^N(B + x) = \mu_L^N(B)$, for every $B \in \mathcal{B}\left(\mathbb{R}^N\right)$ and every $x \in \mathbb{R}^N$;*
3. *$\mu_L^N\left(\mathsf{X}_{K=1}^N (a_K, b_K)\right) = \mathsf{X}_{K=1}^N [a_K, b_K] = \mathsf{X}_{K=1}^N (a_K, b_K] = \mathsf{X}_{K=1}^N [a_K, b_K) = \prod_{K=1}^N (b_K - a_K)$, for all $a_K, b_K \in \mathbb{R}$, such that $a_K \leq b_K$, for every $K = 1, \ldots, N$.*

Definition 57 *We call the function $\mu_L^N : \mathcal{B}\left(\mathbb{R}^N\right) \to \mathbb{R}_+$ presented in Theorem (**??**) the Borel-Lebesgue measure on $\mathbb{R}^N$.*

Let $\mu_L^N : \mathcal{B}\left(\mathbb{R}^N\right) \to \mathbb{R}_+$ be the Borel-Lebesgue measure on $\mathcal{B}\left(\mathbb{R}^N\right)$.

Corollary 58 *We have*

1. *$\mu_L^N(\varnothing) = 0$;*
2. *$\mu_L^N(x) = 0$, for every $x \in \mathbb{R}^N$, where $\mu_L^N(x)$ is the standard shorthand for $\mu_L^N(\{x\})$;*
3. *$\mu_L^N(B) = 0$, for every $B \in \mathcal{B}\left(\mathbb{R}^N\right)$ such that $|B| \leq \aleph_0$.*

Definition 59 *Consider a predicate p on $\mathbb{R}^N$ (see Definition 1289) we say that p holds true* for almost every $x \in \mathbb{R}^N$ *or* almost everywhere *in $\mathbb{R}^N$ if there exist a $N \in \mathcal{B}\left(\mathbb{R}^N\right)$ such that*

$$\mu_L^N(N) = 0 \quad \text{and} \quad \left\{x \in \mathbb{R}^N : p(x) = FALSE\right\} \subseteq N.$$

1.0.5 Lebesgue Integral on $\mathbb{R}^N$

Let $\mathbb{R}^N$ be the N-dimensional real Euclidean space, for some $N \in \mathbb{N}$, let $\mathcal{B}\left(\mathbb{R}^N\right)$ be the Borel σ-algebra of $\mathbb{R}^N$, and let $\mu_L^N : \mathcal{B}\left(\mathbb{R}^N\right) \to \mathbb{R}_+$ be the Borel-Lebesgue measure on $\mathcal{B}\left(\mathbb{R}^N\right)$.

Definition 60 *We say that a function $s : \mathbb{R}^N \to \mathbb{R}$ is a* positive step function *on $\mathbb{R}^N$, if, for some $m \in \mathbb{N}$, there exist $\alpha_1, \ldots, \alpha_m \in \mathbb{R}_+$ and pairwise disjoint $I_1, \ldots, I_m \in \mathcal{I}_{c.o.\mathbb{R}}\left(\mathbb{R}^N\right)$ such that*

$$s(x) = \sum_{j=1}^m \alpha_j 1_{I_j}(x),$$

for every $x \in \mathbb{R}^N$, where $1_{I_j} : \mathbb{R}^N \to \mathbb{R}$ is the indicator function of I_j, for every $j = 1, \ldots, m$.

Lemma 61 *For some $m, n \in \mathbb{N}$, let $\alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_n \in \mathbb{R}_+$, $I_1, \ldots, I_m, J_1, \ldots, J_n \in \mathcal{I}_{c.o.\mathbb{R}}\left(\mathbb{R}^N\right)$ such that*

$$\sum_{j=1}^m \alpha_j 1_{I_j}(x) = \sum_{k=1}^n \beta_k 1_{J_k}(x) \qquad [\text{resp. } \sum_{j=1}^m \alpha_j 1_{I_j}(x) \leq \sum_{k=1}^n \beta_k 1_{J_k}(x)\,]$$

for every $x \in \mathbb{R}^N$, then

$$\sum_{j=1}^m \alpha_j \mu_L^N(I_j) = \sum_{k=1}^n \beta_k \mu_L^N(J_k) \qquad [\text{resp. } \sum_{j=1}^m \alpha_j \mu_L^N(I_j) \leq \sum_{k=1}^n \beta_k \mu_L^N(J_k)\,].$$

Definition 62 *We say that a function $s : \mathbb{R}^N \to \mathbb{R}$ is a* positive simple function*, if, for some $m \in \mathbb{N}$, there exist $\alpha_1, \ldots, \alpha_m \in \mathbb{R}_+$ and pairwise disjoint $A_1, \ldots, A_m \in \mathcal{B}\left(\mathbb{R}^N\right)$ such that*

$$s(x) = \sum_{j=1}^m \alpha_j 1_{A_j}(x) \tag{1.21}$$

for every $x \in \mathbb{R}^N$, where $1_{A_j} : \mathbb{R}^N \to \mathbb{R}$ is the indicator function of A_j, for every $j = 1, \ldots, m$.

Remark 63 *A function $s : \mathbb{R}^N \to \mathbb{R}$ is a positive simple function if and only if it takes a finite number of pasitive values, say $\alpha_1, \ldots, \alpha_m \in \mathbb{R}_+$, for some $m \in \mathbb{N}$. In this case, setting $A_j \equiv \{s^{-1}(\alpha_j)\} \in \mathcal{B}\left(\mathbb{R}^N\right)$, for every $j = 1, \ldots, m$, we can write*

$$s(x) = \sum_{j=1}^m \alpha_j 1_{A_j}(x)\,. \tag{1.22}$$

Definition 64 *If $s : \mathbb{R}^N \to \mathbb{R}$ is a positive simple function, retaining the notation of Remark 63, the right hand side of Equation (1.22) is called the canonical representation of s.*

Lemma 65 *For some $m, n \in \mathbb{N}$, let $\alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_n \in \mathbb{R}_+$ $A_1, \ldots, A_m, B_1, \ldots, B_n \in \mathcal{B}\left(\mathbb{R}^N\right)$ such that*

$$\sum_{j=1}^m \alpha_j 1_{A_j}(x) = \sum_{k=1}^n \alpha_k 1_{B_k}(x) \qquad [\text{resp. } \sum_{j=1}^m \alpha_j 1_{A_j}(x) \leq \sum_{k=1}^n \beta_k 1_{B_k}(x)\,] \tag{1.23}$$

for almost every $x \in \mathbb{R}^N$, then

$$\sum_{j=1}^m \alpha_j \mu_L^N(A_j) = \sum_{k=1}^n \beta_k \mu_L^N(B_k) \qquad [\text{resp. } \sum_{j=1}^m \alpha_j \mu_L^N(A_j) \leq \sum_{k=1}^n \beta_k \mu_L^N(B_k)\,]. \tag{1.24}$$

In Equations (1.23) and (1.24) we use the standard conventions

$$x\mu_L^N(B) = \infty, \quad \forall x \in \mathbb{R}_{++},\ B \in \mathcal{B}\left(\mathbb{R}^N\right)\ \text{s.t. } \mu_L^N(B) = \infty$$

and

$$0\mu_L^N(B) = 0 \quad \forall B \in \mathcal{B}\left(\mathbb{R}^N\right).$$

Definition 66 *We say that a function $f : \mathbb{R}^N \to \mathbb{R}$ is a* positive Borel function*, if we have*

$$f^{-1}(B) \in \mathcal{B}\left(\mathbb{R}^N\right), \quad \forall B \in \mathcal{B}\left(\mathbb{R}^N\right),$$

and

$$f(x) \geq 0,$$

for almost every $x \in \mathbb{R}^N$.

Lemma 67 *For any positive Borel function $f : \mathbb{R}^N \to \mathbb{R}$, there exists $(s_n)_{n=1}^\infty$ sequence of positive simple functions such that*

$$s_n(x) \leq s_{n+1}(x) \quad \text{and} \quad \lim_{n\to\infty} s_n(x) = f(x)$$

almost everywhere in $\mathbb{R}^N$.

Part II

Elements of Descriptive Statistics

Chapter 2

Introduction

A systematic collection of data on the population and the economy was begun in the Italian city-states of Venice and Florence during the Renaissance. the term *statistics*, derived from the word *state*, was used to refer to a collection of facts of interest to the state. Otherwise saying, *statistics* was intended to be a shorthand for the *descriptive science of states*. It was not until the late 1800s that statistics became concerned with inferring conclusions from data. This new perspective began with Francis Galton's work on the study of hereditary genius through the uses of what we would now call regression and correlation analysis.

Statistics has Then, for its object that of presenting a faithful representation of a state at a determined epoch.

Quetele, 1849

Statistics is a scientific discipline concerned with collection, analysis, and interpretation of data obtained from observation or experiment. The subject has a coherent structure based on the theory of Probability and includes many different procedures which contribute to research and development throughout the whole of Science and Technology

E. Pearson, 1936

Statistics is the art of learning from data. It is concerned with the collection of data, their subsequent description, and their analysis, which often leads to the draw of conclusions.

S.H. Ross, 2010

By the term *statistic* is to be intended any numerical quantity determined by a data set.

Sometimes a statistical analysis begins with a given set of data. In other circumstances, data are not available in advance and we can use statistics to design an appropriate experiment to generate data.

The part of statistics concerned with the description and summarization of data is called *descriptive statistics*. The part of statistics concerned with the draw of conclusion from data is called *inferential statistics*.

To draw a conclusion from data, we have to take into account the possibility of chance. Therefore, it is usually necessary to make some assumptions about the chances, to say the probabilities, of obtain the different data values. The totality of these assumptions is referred to as a *probability model* for the data. Statistical inference starts with the assumption that important features of the experiment under consideration can be described in terms of probabilities. Hence, conclusions are drawn from using data to make inferences on these probabilities.

We use the notion of a mathematical model, which we try to fit to the sample data. If this model provides a reasonable fit to the data, that is, if it can approximate the manner in which the data vary, Then, we assume that it can also approximate the behavior of the population.

The model Then, provides the basis for making decisions about the population, by, for example, identifying patterns, explaining variation, and predicting future values. Of course, this process can work only if the sample data can be considered representative of the population.

Since the real world can be extremely complicated (in the way that data values vary or interact together), models are useful because they simplify problems so that we can better understand them (and Then, make more effective decisions). On the one hand, we therefore need models to be simple enough that we can easily use them to make decisions, but on the other hand, we need models that are flexible enough to provide good approximations to complex situations. Fortunately, many statistical models have been developed over the years that provide an effective balance between these two criteria.

Chapter 3

Data Sets

Let $\mathbb{R}^N$ be the real N-dimensional Euclidean space, for some $N \in \mathbb{N}$.

Definition 68 (N-variate real data set) *We call an* N-variate real data set *any sequence* $(x_k)_{k=1}^n \equiv \mathbf{x}$ *of* n *points in* $\mathbb{R}^N$*, where* $n \in \mathbb{N}$*. We call the* length *of the data set* $\mathbf{x}$ *the number* n *of the points in* $\mathbf{x}$*. In case* $N = 1$*, we simply speak of* real data set*.*

A data set aims to give a quantitative or categorical description of a population or a sample drawn from a population. It should be stressed that basic statistics aims to deal with data sets in which the temporal order of data collection is irrelevant, referred to as *cross-sectional* data sets. By shuffling the elements in a cross-sectional data set we are supposed to build another data set with the same statistical properties. When the temporal order of data collection is important the data set should be more properly called a *time series*. It may be interesting to compare the two notions.

Definition 69 (N-variate real time series) *We call an* N-variate real time series *any sequence* $(x_t)_{t \in T} \equiv \mathbf{x}$ *of* n *points in* $\mathbb{R}^N$*, where* $T \subseteq \mathbb{R}$ *is a finite set of time indices. We call the* length *of the time series* $\mathbf{x}$*, denoted by* $|T|$*, the number of time indices in* T*. In case* $N = 1$*, we simply speak of* real time series*.*

Differently than a cross-sectional data set, the temporal ordering of the elements of a time series, given by the time indices, plays a crucial role in characterizing time series, even when it is possible to prove its lack of influence: if shuffling the elements of a time series, we built another time series with the same statistical properties, we would have discovered an essential property of the time series.

Convention 70 *From now on, we adopt the standard convention that when algebraically manipulating* $N_1, \ldots, N_p$*-variate real data sets of various lengths* $n_1, \ldots, n_p$*, respectively, for some* $p \in \mathbb{N}$*, let us denote them* $\mathbf{x}_1 \equiv (x_k)_{k=1}^{n_1}, \ldots, \mathbf{x}_p \equiv (x_k)_{k=1}^{n_p}$*, the elements of each data set* $\mathbf{x}_\ell \equiv (x_k)_{k=1}^{n_\ell}$*, will be thought of as column vectors in the real Euclidean spaces,* $\mathbb{R}^{N_\ell}$*, for* $\ell = 1, \ldots, p$*.. The data set* $\mathbf{x}_\ell$ *itself will be thought of as a matrix in the real Euclidean spaces*

$\mathbb{R}^{N_\ell \times n_\ell}$, *for* $\ell = 1, \ldots, p$. *In symbols,*

$$\mathbf{x}_1 \equiv (x_k)_{k=1}^{n_1} \equiv (x_1, \ldots, x_{n_1}) \equiv \begin{pmatrix} x_{1,1} & \cdots & x_{1,n_1} \\ \vdots & \ddots & \vdots \\ x_{N_1,1} & \cdots & x_{N_1,n_1} \end{pmatrix} \in \mathbb{R}^{N_1 \times n_1},$$

$$\ldots$$

$$\mathbf{x}_p \equiv (x_k)_{k=1}^{n_p} \equiv \left(x_1, \ldots, x_{n_p}\right) \equiv \begin{pmatrix} x_{1,1} & \cdots & x_{1,n_p} \\ \vdots & \ddots & \vdots \\ x_{N_p,1} & \cdots & x_{N_p,n_p} \end{pmatrix} \in \mathbb{R}^{N_p \times n_p},$$

In particular, when $N = 1$, *a real data set of length* n, *will be thought of as a row vector in the Euclidean space* $\mathbb{R}^n$. *That is*

$$(x_k)_{k=1}^{n} \equiv (x_1, \ldots, x_n) \in \mathbb{R}^n.$$

3.1 Statistics on Data Sets: Mode, Mean

Definition 71 *We call an* M-variate real statistics *computed on the data set* $\mathbf{x}$ *the value* $g(\mathbf{x})$ *of any Borel function* $g : \mathbb{R}^N \to \mathbb{R}^M$, *where* $M \in \mathbb{N}$. *In case* $N = 1$,*we briefly speak of* real statistics.

Let $(x_k)_{k=1}^{n} \equiv \mathbf{x}$ be an N-variate real data set of length n.

Definition 72 *We call the* set of values *of the data set* $\mathbf{x}$ *the subset of* $\mathbb{R}^N$

$$\mathbf{x}_{\{\}} \stackrel{def}{=} \left\{x \in \mathbb{R}^N : x = x_k, \quad k = 1, \ldots, n\right\}.$$

Remark 73 *We clearly have*

$$1 \leq \left|\mathbf{x}_{\{\}}\right| \leq n,$$

where $\left|\mathbf{x}_{\{\}}\right|$ *denotes the number of the elements in* $\mathbf{x}_{\{\}}$.

Notation 74 *When we need to refer to the set of values of a data set as a data set itself, we write* $\mathbf{x}_{\{\}} \equiv \{x_j\}_{j=1}^{m}$, *where* $m \equiv \left|\mathbf{x}_{\{\}}\right|$.

Let $\mathbf{x}_{\{\}} \equiv \{x_j\}_{j=1}^{m}$ be the set of values of the data set $(x_k)_{k=1}^{n} \equiv \mathbf{x}$.

Definition 75 *We call the* frequency *of the value* x_j *in the data set* $\mathbf{x}$ *the positive integer* f_j *defined by*

$$f_j \stackrel{def}{=} \left|\{k \in \{1, \ldots, n\} : x_k = x_j\}\right|, \quad \forall j = 1, \ldots, m.$$

We call the relative frequency *of the value* x_j *in the data set* $\mathbf{x}$ *the positive rational number* r_j *defined by*

$$r_j \stackrel{def}{=} \frac{f_j}{n}, \quad \forall j = 1, \ldots, m,$$

where f_j *is the frequency of the value* x_j *in* $\mathbf{x}$.

Remark 76 *We clearly have*

$$\sum_{j=1}^{m} f_j = n \quad \textit{and} \quad \sum_{j=1}^{m} r_j = 1.$$

Definition 77 *The data set* $\mathbf{x}$ *is said to be* symmetric about the value $x_0 \in \mathbb{R}^N$ *if considering the data set* $\left(x_k^{(0)}\right)_{k=1}^n \equiv \mathbf{x}^{(0)}$ *given by*

$$x_k^{(0)} \stackrel{def}{=} x_k - x_0, \quad \forall k = 1, \ldots, n,$$

the set of values $\left\{x_j^{(0)}\right\}_{j=1}^m$ *of* $\mathbf{x}^{(0)}$ *is symmetric about* 0 *and the frequencies* f_j *[resp. relative frequencies* r_j*] of the symmetrically corresponding values are the same for every* $j = 1, \ldots, m$.

Definition 78 *We call* a mode *of the data set* $\mathbf{x}$ *any* $x_{mo} \in \{x_j\}_{j=1}^m$ *given by*

$$x_{mo} \stackrel{def}{=} x_{\check{j}}, \quad \check{j} \equiv \arg\max_{j=1,\ldots,m} \{f_j\}. \tag{3.1}$$

where f_j *is the frequency of* x_j *in* $\mathbf{x}$*. In case* x_{mo} *is unique, we speak of* the mode *of* x.

In words, a mode of a data set is any value which occurs the most frequently.

Definition 79 *We call the* mean *of the data set* $\mathbf{x}$ *the real number defined by*

$$\bar{x}_n \stackrel{def}{=} \frac{1}{n}\sum_{k=1}^n x_k. \tag{3.2}$$

Remark 80 *We have*

$$\bar{x}_n = \frac{1}{n}\sum_{j=1}^m f_j x_j = \sum_{j=1}^m r_j x_j.$$

where f_j *[resp.* r_j*] is the frequency [resp. relative frequency] of the value* x_j *in* $\mathbf{x}_{\{\}}$*, for every* $j = 1, \ldots, m$.

Exercise 81 *We consider a population of* $n = 40$ *individuals and for each individual in the sample, labeled by the index* $k = 1, \ldots, n$*, we check the height and weight expressed in meters and kilograms respectively. As a result, we obtain the following table*

k	*height*	*weight*	k	*height*	*weight*	k	*height*	*weight*	k	*height*	*weight*
1	1.72	78.600	11	1.76	79.300	21	1.77	79.400	31	1.56	62.500
2	1.73	78.900	12	1.64	63.400	22	1.65	63.400	32	1.52	62.400
3	1.71	64.000	13	1.53	62.300	23	1.68	63.600	33	1.71	78.100
4	1.73	78.300	14	1.92	80.500	24	1.73	80.100	34	1.53	62.100
5	1.74	77.700	15	1.61	63.300	25	1.67	77.800	35	1.69	63.700
6	1.81	79.100	16	1.70	79.200	26	1.44	61.200	36	1.52	62.100
7	1.59	62.800	17	1.79	77.800	27	1.66	77.700	37	1.49	61.700
8	1.80	79.200	18	1.81	79.100	28	1.72	63.900	38	1.73	79.000
9	1.79	79.400	19	1.63	79.100	29	1.75	64.300	39	1.57	62.700
10	1.54	62.300	20	1.73	78.500	30	1.55	62.300	40	1.62	63.100

(3.3)

With reference to data sets $(x_k)_{k=1}^n \equiv \mathbf{x}$ *presented in Table 3.3 determine the set of values, and compute the frequencies, relative frequencies, mode, and mean. Compute also such statistics for both the two entries of* $\mathbf{x}$ *separately.*

Let $(x_k)_{k=1}^n \equiv \mathbf{x}$ [resp. $(y_k)_{k=1}^n \equiv \mathbf{y}$] be an M-variate [resp. N-variate] real data set of length n.

Proposition 82 *Assume*

$$y_k \overset{def}{=} Ax_k + \beta, \quad \forall k = 1, \dots, n,$$

for some $A \in \mathbb{R}^{N\times M}$ *and* $\beta \in \mathbb{R}^N$. *Then, we have*

$$\bar{y}_n = A\bar{x}_n + \beta,$$

where $\bar{x}_n$ *[resp.* $\bar{y}_n$*] is the mean of the data sets* $\mathbf{x}$ *[resp.* $\mathbf{y}$*].*

Proof. . □

3.2 Statistics on Data Sets: Order Statistics, Median, Quantiles

Set $N = 1$ and let $(x_k)_{k=1}^n \equiv \mathbf{x}$ be a real data set of length n.

Definition 83 *We call the* order statistics *of the data set* $\mathbf{x}$*, and we denote it by* $\left(x_{(k)}\right)_{k=1}^n$ *or briefly by* $\mathbf{x}_{()}$*, any non-decreasing reordering of* $\mathbf{x}$.

Let $\left(x_{(k)}\right)_{k=1}^n \equiv \mathbf{x}_{()}$ be the order statistics of the data set $(x_k)_{k=1}^n \equiv \mathbf{x}$.

Definition 84 *For any* $k = 1, \dots, n$*, we call the real number* $x_{(k)}$ *the* k*th* order statistic *of* $\mathbf{x}$. *In particular,*

$$x_{(1)} = \min\{x_1, \dots, x_n\} \quad \textit{and} \quad x_{(n)} = \max\{x_1, \dots, x_n\}$$

are the 1*st and* n*th order statistic, respectively.*

Definition 85 *We call the* range *of* $\mathbf{x}$ *the real number*

$$\text{range}\,(\mathbf{x}) \overset{def}{=} x_{(n)} - x_{(1)}. \tag{3.4}$$

Definition 86 *We call the* median *of* $\mathbf{x}$ *the real number*

$$x_{1/2} \overset{def}{=} \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) & \text{if } n \text{ is even} \end{cases}. \tag{3.5}$$

In light of (3.5), when the length n of the data set $\mathbf{x}$ is odd [resp. even] the median $x_{1/2}$ is the middle value [resp. the average of the two middle values] of $\mathbf{x}_{()}$. Hence, when n is odd the median $x_{1/2}$ is always in the set of values $\mathbf{x}_{\{\}}$ of $\mathbf{x}$. Instead, when n is even $x_{1/2}$ may not be in $\mathbf{x}_{\{\}}$. Note that the median, which makes use of only one or two middle values of the time series $\mathbf{x}$, is not affected by the extreme values of $\mathbf{x}$. On the contrary, the mean of $\mathbf{x}$, which makes use of the whole data set, is affected by the extreme values. Therefore, the median is a more accurate measure of the centrality of $\mathbf{x}$ than the mean according to whether the extreme values of $\mathbf{x}$ matter. However, when $\mathbf{x}$ is roughly symmetric about the mean, the median attains a value close to the mean.

Although it is interesting to wonder whether the mean or median is more useful in a particular situation, we never need to restrict ourselves to consider only one of these statistics. They are both important, and should always be computed when a data set is summarized.

Let $\lfloor\cdot\rfloor : \mathbb{R} \to \mathbb{Z}$ the *floor* or *integer part* function (see Chapter 1).

Definition 87 *Given any integer m such that $0 \leq m < 50$, we call the $m\%$ trimmed mean of* $\mathbf{x}$ *the real number*

$$\text{mean}_{\mathbf{tr}(m)}(\mathbf{x}) \stackrel{def}{=} \frac{1}{n - \lfloor n \cdot 2m/100 \rfloor} \sum_{k=\lfloor n \cdot m/100 \rfloor + 1}^{n - \lfloor n \cdot m/100 \rfloor} x_{(k)}, \tag{3.6}$$

where $x_{(k)}$ is the kth order statistic of $\mathbf{x}$ *for any $k = \lfloor n \cdot m/100 \rfloor + 1, \ldots, n - \lfloor n \cdot m/100 \rfloor$.*

Exercise 88 *With reference to Exercise 81, compute the median and the 10% trimmed mean of both the height and weight entries of the data set* $\mathbf{x}$ *separately.*

Definition 89 (Sheldon M. Ross) *Given any rational number $q \in (0, 1)$, set $h \equiv (n+1)\, q$. We call the* quantile of order q *or* q-quantile *or* $100qth$ percentile *of the data set* $\mathbf{x}$ *the real number*

$$x_q \stackrel{def}{=} x_{(\lfloor h \rfloor)} + (h - \lfloor h \rfloor) \left(x_{(\lfloor h \rfloor + 1)} - x_{(\lfloor h \rfloor)} \right). \tag{3.7}$$

In particular,

- *the 0.25-quantile or 25th percentile is also called the* first quartile *and is denoted by x_{q_1};*
- *the 0.50-quantile or 50th percentile is also called the* second quartile*, it coincides with the median, and it is also denoted by x_{q_2};*
- *the 0.75-quantile or 75th percentile is also called the* third quartile *and is denoted by x_{q_3}.*

Remark 90 *Setting $q \equiv k/(n+1)$, we have*

$$x_q = x_{(k)},$$

for every $k = 1, \ldots, n$.

Remark 91 *For any $q \in (0, 1)$, the q-quantile x_q has at least $\lfloor nq \rfloor$ data [resp. $\lfloor n(1-q) \rfloor$] of the order statistic less [resp. greater] than it.*

In particular, the quartiles break up a data set into four parts. About the 25% of the data values are smaller than the first quartile, about the 25% of the data values lie between the first and second quartile, about the 25% are between the second and third quartile, and about the 25% are larger than the third quartile.

Definition 92 *We call the* interquartile range *of* $\mathbf{x}$*, acronym IQR, the difference between the first and third quartile. In symbols,*

$$IQR = x_{q_3} - x_{q_1}.$$

Example 93 *Consider the following list of the average Celsius temperatures recorded in Roma during the first half of September* 2017

$$\begin{pmatrix} 25.2 & 25.2 & 21.8 & 23.9 & 25.6 & 24.9 & 25.2 & 24.8 & 23.8 & 19.5 & 20.3 & 22.2 & 23.8 & 22.9 & 25.3 \end{pmatrix}$$

Determine the quartiles and some percentiles.

Discussion. To compute percentiles we have first to consider the order statistic of length $n = 15$, that is

$$\begin{pmatrix} 19.5 & 20.3 & 21.8 & 22.2 & 22.9 & 23.8 & 23.8 & 23.9 & 24.8 & 24.9 & 25.2 & 25.2 & 25.2 & 25.3 & 25.6 \end{pmatrix}.$$

Now, setting $q = 1/2$, we have that $(n+1)\,q = 8 = \lfloor (n+1)\,q \rfloor$. Hence, the 50th percentile or second quartile or median is given by the 8th element of the order statistic, that is

$$x_{q_2} \equiv x_{1/2} = x_{(8)} \equiv 23.9.$$

Furthermore, $\lfloor nq \rfloor = \lfloor 15/2 \rfloor = 7$ and $\lfloor n\,(1-q) \rfloor = \lfloor 15/2 \rfloor = 7$ In fact, there are 7 elements of the order statistic which are less than 23.9 and 7 elements which are greater than 23.9.

Setting $q = 1/4$, we have that $(n+1)\,q = 4 = \lfloor (n+1)\,q \rfloor$. Hence, the 25th percentile or first quartile is given by the 4th element of the order statistic, that is

$$x_{q_1} = x_{(4)} \equiv 22.2.$$

Furthermore, $\lfloor nq \rfloor = \lfloor 15/4 \rfloor = 3$ and $\lfloor n\,(1-q) \rfloor = \lfloor 45/4 \rfloor = 11$. In fact, there are 3 elements of the order statistic which are less than 23.9 and 11 elements which are greater than 23.9.

Setting $q = 1/10$, we have that $(n+1)\,q = 16/10 > \lfloor (n+1)\,q \rfloor = 1$ and $(n+1)\,q - \lfloor (n+1)\,q \rfloor = 6/10$. Hence, the 10th percentile is given by

$$x_{0.1} = x_{(1)} + \frac{6}{10}\left(x_{(2)} - x_{(1)}\right) = 19.5 + 0.6\,(20.3 - 19.5) = 19.98 \simeq 20.$$

Furthermore, $\lfloor nq \rfloor = \lfloor 15/10 \rfloor = 1$ and $\lfloor n\,(1-q) \rfloor = \lfloor 135/10 \rfloor = 13$. In fact, there is 1 element of the order statistic which is less than 20 and 14 elements which are greater than 22.

Setting $q = 1/(n+1) = 0.0625$, we have that $(n+1)\,q = 1 = \lfloor (n+1)\,q \rfloor$. Hence, the 6.25th percentile is given by the 1th element of the order statistic, that is

$$x_{1/16} = x_{(1)} = 19.5.$$

Furthermore, $\lfloor nq \rfloor = \lfloor n/\,(n+1) \rfloor = \lfloor 15/16 \rfloor = 0$ and $\lfloor n\,(1-q) \rfloor = \lfloor n\,(n/\,(n+1)) \rfloor = \lfloor 15^2/16 \rfloor = 14$. In fact, there are 0 elements of the order statistic which are less than 19.5 and 14 elements which are greater than 19.5. □

Exercise 94 *With reference to Example 93, consider the following list of the average Celsius temperatures recorded in Roma during the second half of September* 2017

$$\begin{pmatrix} 25.7 & 22.0 & 21.2 & 20.7 & 18.3 & 19.5 & 20.7 & 20.8 & 20.6 & 21.7 & 20.5 & 21.5 & 22.2 & 21.4 & 20.6 \end{pmatrix}$$

Determine the quartiles, the IQR, and some further percentiles of the average Celsius temperatures recorded in Roma throughout September 2017*.*

Note that Definition 89 is just one of the various not completely equivalent definitions that it is possible to find in the statistical literature. For instance, another definition is the following

Definition 95 (Jay L. Devore & Kenneth N. Berk) *Given any $k = 1, \ldots, n$, we call the $100\,(k-0.5)\,/n$th* percentile *or $(k-0.5)\,/n$* quantile *of the time series $\mathbf{x}$ the real number $x_{(k)}$ which occupies the kth place in the order statistics $\mathbf{x}_{()}$ of $\mathbf{x}$.*

Definition 89 is focused on determining the quantile x_q of the data set $\mathbf{x}$ corresponding to any $q \in (0,1)$. In contrast, Definition 95 is focused on determining the quantile order q corresponding to each point $x_{(k)}$ of the order statistics $\mathbf{x}_{()}$ of $\mathbf{x}$.

Exercise 96 *With reference to Exercise 81, determine the order statistics and compute the quartiles for both the height and weight entries of the data set $\mathbf{x}$.*

3.3 Statistics on Data Sets: Deviation, Variance-covariance, Correlation, Standard Deviation

Let $(x_k)_{k=1}^n \equiv \mathbf{x}$ be an N-variate real data set of length n, and let $\bar{x}_n$ be the mean of $\mathbf{x}$.

Definition 97 *For any $k = 1, \ldots, n$, we call* deviation *of x_k from $\bar{x}_n$ the difference*

$$x_k - \bar{x}_n. \tag{3.8}$$

Remark 98 *We clearly have*

$$\sum_{k=1}^n (x_k - \bar{x}_n) = 0. \tag{3.9}$$

Definition 99 *We call the* variance-covariance *of $\mathbf{x}$ the matrix in $\mathbb{R}^{N\times N}$*

$$s^2_{\mathbf{x},n} \overset{def}{=} \frac{1}{n}\sum_{k=1}^n (x_k - \bar{x}_n)(x_k - \bar{x}_n)^\intercal, \tag{3.10}$$

where $\intercal$ is the transpose operator. *Note that in case $N = 1$ the variance-covariance of $\mathbf{x}$ reduces to the real number*

$$s^2_{\mathbf{x},n} \overset{def}{=} \frac{1}{n}\sum_{k=1}^n (x_k - \bar{x}_n)^2 \tag{3.11}$$

*and it is briefly called the (*biased) variance *of $\mathbf{x}$.*

Remark 100 *We have*

$$s^2_{\mathbf{x},n} = \frac{1}{n}\left(\sum_{k=1}^n x_k x_k^\intercal - n\bar{x}_n\bar{x}_n^\intercal\right). \tag{3.12}$$

In particular, in case $N = 1$, we obtain

$$s^2_{\mathbf{x},n} = \frac{1}{n}\left(\sum_{k=1}^n x_k^2 - n\bar{x}_n^2\right). \tag{3.13}$$

Proof. By direct computation

$$\begin{aligned}
s^2_{\mathbf{x},n} &= \frac{1}{n}\sum_{k=1}^n (x_k - \bar{x}_n)(x_k - \bar{x}_n)^\intercal = \frac{1}{n}\sum_{k=1}^n (x_k - \bar{x}_n)(x_k^\intercal - \bar{x}_n^\intercal) \\
&= \frac{1}{n}\sum_{k=1}^n x_k x_k^\intercal - \frac{1}{n}\sum_{k=1}^n x_k\bar{x}_n^\intercal - \frac{1}{n}\sum_{k=1}^n \bar{x}_n x_k^\intercal + \frac{1}{n}\sum_{k=1}^n \bar{x}_n\bar{x}_n^\intercal \\
&= \frac{1}{n}\sum_{k=1}^n x_k x_k^\intercal - \left(\frac{1}{n}\sum_{k=1}^n x_k\right)\bar{x}_n^\intercal - \bar{x}_n\left(\frac{1}{n}\sum_{k=1}^n x_k^\intercal\right) + \bar{x}_n^\intercal\bar{x}_n^\intercal \\
&= \frac{1}{n}\sum_{k=1}^n x_k x_k^\intercal - \bar{x}_n\bar{x}_n^\intercal - \bar{x}_n\bar{x}_n^\intercal + \bar{x}_n^\intercal\bar{x}_n^\intercal \\
&= \frac{1}{n}\left(\sum_{k=1}^n x_k x_k^\intercal - n\bar{x}_n\bar{x}_n^\intercal,\right),
\end{aligned}$$

as desired. □

Proposition 101 *The variance covariance matrix $s^2_{\mathbf{x},n}$ is symmetric and positive semidefinite.*

Proof. By direct computation, considering the properties of the transpose operator, we have

$$\begin{aligned}\left(s^2_{\mathbf{x},n}\right)^\intercal &= \left(\frac{1}{n}\sum_{k=1}^n (x_k-\bar{x}_n)(x_k-\bar{x}_n)^\intercal\right)^\intercal \\ &= \frac{1}{n}\sum_{k=1}^n \left((x_k-\bar{x}_n)(x_k-\bar{x}_n)^\intercal\right)^\intercal \\ &= \frac{1}{n}\sum_{k=1}^n \left((x_k-\bar{x}_n)^\intercal\right)^\intercal (x_k-\bar{x}_n)^\intercal \\ &= \frac{1}{n}\sum_{k=1}^n (x_k-\bar{x}_n)(x_k-\bar{x}_n)^\intercal \\ &= s^2_{\mathbf{x},n}.\end{aligned}$$

This shows that $s^2_{\mathbf{x},n}$ is symmetric. Now, since for any $u \in \mathbb{R}^N$ we have $u^\intercal (x_k-\bar{x}_n) \in \mathbb{R}$, we can write

$$\begin{aligned}u^\intercal s^2_{\mathbf{x},n} u &= \frac{1}{n} u^\intercal \left(\sum_{k=1}^n (x_k-\bar{x}_n)(x_k-\bar{x}_n)^\intercal\right) u \\ &= \frac{1}{n}\sum_{k=1}^n u^\intercal (x_k-\bar{x}_n)(x_k-\bar{x}_n)^\intercal u \\ &= \frac{1}{n}\sum_{k=1}^n (x_k-\bar{x}_n)^\intercal u\,(x_k-\bar{x}_n)^\intercal u \\ &= \frac{1}{n}\sum_{k=1}^n \left((x_k-\bar{x}_n)^\intercal u\right)^2.\end{aligned}$$

It clearly follows that

$$u^\intercal s^2_{\mathbf{x},n} u \geq 0$$

for every $u \in \mathbb{R}^N$, which is the positive semidefiniteness of $s^2_{\mathbf{x},n}$. □

Proposition 102 *We have $s^2_{\mathbf{x},n} = 0$ if and only if $x_1 = x_2 = \cdots = x_n$.*

Proof. It is clearly seen that $x_1 = x_2 = \cdots = x_n \equiv x_0$ implies $\bar{x}_n = x_0$. It follows

$$\sum_{k=1}^n (x_k-\bar{x}_n)(x_k-\bar{x}_n)^\intercal = 0,$$

that is

$$s^2_{\mathbf{x},n} = 0.$$

To prove the converse, observe that, in terms of the entries of the vector $x_k - \bar{x}_n$ (see Convention 70), we have

$$(x_k-\bar{x}_n)(x_k-\bar{x}_n)^\intercal = \begin{pmatrix} x_{1,k}-\bar{x}_{1,n} \\ x_{2,k}-\bar{x}_{2,n} \\ \vdots \\ x_{N-1,k}-\bar{x}_{N-1,,n} \\ x_{N,k}-\bar{x}_{N,,n}\end{pmatrix} \begin{pmatrix} x_{1,k}-\bar{x}_{1,n} & x_{2,k}-\bar{x}_{2,n} & \cdots & x_{N-1,k}-\bar{x}_{N-1,,n} & x_{N,k}-\bar{x}_{N,,n}\end{pmatrix}$$

$$= \begin{pmatrix} (x_{1,k}-\bar{x}_{1,n})^2 & (x_{1,k}-\bar{x}_{1,n})(x_{2,k}-\bar{x}_{2,n}) & \cdots & (x_{1,k}-\bar{x}_{1,n})(x_{N-1,k}-\bar{x}_{N-1,,n}) \\ (x_{2,k}-\bar{x}_{2,n})(x_{1,k}-\bar{x}_{1,n}) & (x_{2,k}-\bar{x}_{2,n})^2 & \cdots & (x_{2,k}-\bar{x}_{2,n})(x_{N-1,k}-\bar{x}_{N-1,,n}) \\ \vdots & \vdots & \ddots & \vdots \\ (x_{N-1,k}-\bar{x}_{N-1,,n})(x_{1,k}-\bar{x}_{1,n}) & (x_{N-1,k}-\bar{x}_{N-1,,n})(x_{2,k}-\bar{x}_{2,n}) & \cdots & (x_{N-1,k}-\bar{x}_{N-1,,n})^2 \\ (x_{N,k}-\bar{x}_{N,,n})(x_{1,k}-\bar{x}_{1,n}) & (x_{N,k}-\bar{x}_{N,,n})(x_{2,k}-\bar{x}_{2,n}) & \cdots & (x_{N,k}-\bar{x}_{N,,n})(x_{N-1,k}-\bar{x}_{N-1,,n}\end{pmatrix}$$

for every $k = 1, \ldots, n$. Therefore, the diagonal entries of the matrix $s^2_{\mathbf{x},n}$ are given by

$$s^2_{\mathbf{x},n,K,K} = \sum_{k=1}^n (x_{K,k} - \bar{x}_{K,n})^2$$

for every $K = 1, \ldots, N$. Then, the assumption $s^2_{\mathbf{x},n} = 0$ implies

$$\sum_{k=1}^n (x_{K,k} - \bar{x}_{K,n})^2 = 0$$

for every $K = 1, \ldots, N$. In turn, the latter implies

$$x_{K,k} = \bar{x}_{K,n}$$

for every $k = 1, \ldots, n$. It follows that

$$x_1 = \cdots = x_n = \bar{x}_n,$$

as desired. □

Corollary 103 *There exists a unique symmetric and positive semidefinite matrix $s_{\mathbf{x},n}$ such that*

$$s_{\mathbf{x},n} s_{\mathbf{x},n} = s^2_{\mathbf{x},n}. \tag{3.14}$$

Definition 104 *We call the* standard deviation *of the data set $\mathbf{x}$ the positive semidefinite matrix $s_{\mathbf{x},n}$ which satisfies Equation (3.14). With algebraic language $s_{\mathbf{x},n}$ is called the* square root *of $s^2_{\mathbf{x},n}$. In case $N = 1$ we have*

$$s_{\mathbf{x},n} = \sqrt{s^2_{\mathbf{x},n}}. \tag{3.15}$$

Definition 105 *Assume* $\det\left(\operatorname{diag}\left(s^2_{\mathbf{x},n}\right)\right) > 0$, *where* $\operatorname{diag}\left(s^2_{\mathbf{x},n}\right)$ *is the diagonal matrix having for diagonal entries the corresponding diagonal entries of $s^2_{\mathbf{x},n}$ and* $\det : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ *is the determinant function on $\mathbb{R}^N \times \mathbb{R}^N$. We call the* correlation *of $\mathbf{x}$ the matrix in $\mathbb{R}^{N\times N}$*

$$r_{\mathbf{x},n} \overset{def}{=} \operatorname{diag}\left(s^2_{\mathbf{x},n}\right)^{-\frac{1}{2}} s^2_{\mathbf{x},n} \operatorname{diag}\left(s^2_{\mathbf{x},n}\right)^{-\frac{1}{2}} \equiv s^{-1}_{\mathbf{x},n} s^2_{\mathbf{x},n} s^{-1}_{\mathbf{x},n}. \tag{3.16}$$

Note that in case $N = 1$ the correlation of $\mathbf{x}$ reduces to the real number 1.

Exercise 106 *With reference to Exercise 81, compute the variance-covariance matrix and the standard deviation of the data set $\mathbf{x}$ and of both the height and weight entries of $\mathbf{x}$ separately.*

Let $(x_k)_{k=1}^n \equiv \mathbf{x}$ [resp. $(y_k)_{k=1}^n \equiv \mathbf{y}$] be an M-variate [resp. N-variate] real data set of length n.

Proposition 107 *Assume*

$$y_k \overset{def}{=} \alpha x_k + \beta, \quad \forall k = 1, \ldots, n,$$

for some $\alpha \in \mathbb{R}^{N\times M}$ and $\beta \in \mathbb{R}^N$. Then, we have

$$s^2_{\mathbf{y},n} = \alpha s^2_{\mathbf{x},n} \alpha^\intercal,$$

where $s^2_{\mathbf{x},n}$ [resp. $s^2_{\mathbf{y},n}$] is the variance-covariance matrix of the data sets $\mathbf{x}$ [resp. $\mathbf{y}$].

Proof. . □

Remark 108 *In case $N = 1$, we have*

$$s^2_{\mathbf{y},n} = \alpha^2 s^2_{\mathbf{x},n}$$

and also

$$s_{\mathbf{y},n} = |\alpha|\, s_{\mathbf{x},n}$$

where $s_{\mathbf{x},n}$ [resp. $s_{\mathbf{y},n}$] is the standard deviation of the data sets $\mathbf{x}$ [resp. $\mathbf{y}$].

Theorem 109 (Čebyšëv Inequality for Data Sets) *In case $N = 1$, fixed any $m > 0$, write*

$$\Omega_m = \{k \in \{1, \ldots, n\} : |x_k - \bar{x}_n| \leq m s_{\mathbf{x},n}\},$$

where $s_{\mathbf{x},n}$ is the standard deviation of the data set $\mathbf{x}$. Then,

$$\frac{|\Omega_m|}{n} > 1 - \frac{1}{m^2}, \tag{3.17}$$

where $|\Omega_m|$ is the cardinality of the set Ω_m.

Proof. By virtue of Equation (3.15) and Remark 102, the Chebyshev inequality is trivially true when $s_{\mathbf{x},n} = 0$. Therefore, we prove it under the further assumption $s_{\mathbf{x},n} > 0$. We clearly have

$$\begin{aligned} n s^2_{\mathbf{x},n} &= \sum_{k=1}^{n} (x_k - \bar{x}_n)^2 = \sum_{k \in \Omega_m} (x_k - \bar{x}_n)^2 + \sum_{k \notin \Omega_m} (x_k - \bar{x}_n)^2 \\ &\geq \sum_{k \notin \Omega_m} (x_k - \bar{x}_n)^2 > \sum_{k \notin \Omega_m} m^2 s^2_{\mathbf{x},n} = (n - |\Omega_m|)\, m^2 s^2_{\mathbf{x},n}. \end{aligned} \tag{3.18}$$

Hence, dividing both the sides of Inequality (3.18) by $n m^2 s^2_{\mathbf{x},n}$, we obtain

$$\frac{1}{m^2} > 1 - \frac{|\Omega_m|}{n},$$

which clearly inplies Equation (3.17). □

Corollary 110 (Čebyšëv Inequality for Data Sets) *With reference to Theorem 109, writing*

$$\Omega^c_m = \{k \in \{1, \ldots, n\} : |x_k - \bar{x}_n| > m s_{\mathbf{x},n}\},$$

we have

$$\frac{|\Omega^c_m|}{n} < \frac{1}{m^2}. \tag{3.19}$$

Proof. We have

$$|\Omega^c_m| = n - |\Omega_m|.$$

Hence, applying Equation (3.17), we obtain

$$|\Omega^c_m| < n - n\left(1 - \frac{1}{m^2}\right) = \frac{n}{m^2}.$$

The desired (3.19) immediately follows. □

Theorem 109 and Corollary 110 enhance the role of the standard deviation $s_{\mathbf{x},n}$ of the data set $\mathbf{x}$ as a measure of dispersion around the mean $\bar{x}_n$. In fact $|\Omega_m|/n$ [resp. $|\Omega_m^c|/n$] is the percentage of points of $\mathbf{x}$ having a distance from $\bar{x}_n$ smaller [resp. larger] than $ms_{\mathbf{x},n}$. According to Theorem 109 [resp. Corollary 110], this percentage is bounded from below [resp. above] by $1-1/m^2$ [resp. $1/m^2$], which does not depend on $s_{\mathbf{x},n}$. This means that whether $s_{\mathbf{x},n}$ is large or small the cardinality of $|\Omega_m|$ [resp. $|\Omega_m^c|$] has the same lower [resp. upper] bound. Threrefore, the number of points in the data set having a distance from $\bar{x}_n$ smaller [resp. larger] than $ms_{\mathbf{x},n}$ is roughly constant. In turn, this implies that the smaller [resp. larger] $s_{\mathbf{x},n}$ the closer [resp. farther] the points in the data set $\mathbf{x}$ are to [resp. from] $\bar{x}_n$.

Example 111 (Čebyšëv Inequality for Data Sets) *With reference to Theorem 109, setting $m = 3$, we obtain that the percentage of the points in a data set $\mathbf{x}$ which are closer to [resp. farther from] the median $\bar{x}_n$ than 3 times the standard deviation $s_{\mathbf{x},n}$ is more than 89% [resp. less than 11 %].*

3.4 Statistics on Data Sets: Moments

Set $N = 1$ and consider a real data set $(x_k)_{k=1}^n \equiv \mathbf{x}$ of length n and fix any $p \in \mathbb{N}$.

Definition 112 *We call the pth* raw moment *of the data set x the real number*

$$\mu'_{\mathbf{x},n,p} \stackrel{def}{=} \frac{1}{n}\sum_{k=1}^n x_k^p. \tag{3.20}$$

Note that $\mu'_{\mathbf{x},n,1} = \bar{x}_n$.

Definition 113 *We call the pth* central moment *of the data set x the real number*

$$\mu_{\mathbf{x},n,p} \stackrel{def}{=} \frac{1}{n}\sum_{k=1}^n (x_k - \bar{x}_n)^p. \tag{3.21}$$

Note that $\mu_{\mathbf{x},n,1} = 0$ and $\mu_{\mathbf{x},n,2} = \tilde{s}^2_{\mathbf{x},n}$.

Definition 114 *Assume $p \geq 2$, we call the pth* standardized central moment *of the data set x the real number*

$$\hat{\mu}_{\mathbf{x},n,p} \stackrel{def}{=} \frac{\mu_{\mathbf{x},n,p}}{s^p_{\mathbf{x},n}}, \tag{3.22}$$

where $s^p_{\mathbf{x},n} \equiv (s_{\mathbf{x},n})^p$ and $s_{\mathbf{x},n} \equiv \sqrt{s^2_{\mathbf{x},n}}$.

Note that $\hat{\mu}_{\mathbf{x},n,2} = 1$.

Definition 115 *The the 3rd [resp. 4th]* standardized central moment *of the data set $\mathbf{x}$ is also known as the* skewness *[resp.* kurtosis*] of $\mathbf{x}$ and it is denoted by the symbol $skew_{\mathbf{x},n}$ [resp. $kurt_{\mathbf{x},n}$].*

3.5 Cross-covariance, Cross-correlation

Let $(x_k)_{k=1}^n \equiv \mathbf{x}$ [resp. $(y_k)_{k=1}^n \equiv \mathbf{y}$] be an M-variate [resp. N-variate] real data set of length n.

Definition 116 *We call the* cross-covariance *of the data sets* $\mathbf{x}$ *and* $\mathbf{y}$ *the matrix in* $\mathbb{R}^{M\times N}$ *given by*

$$s_{\mathbf{x},\mathbf{y},n} \stackrel{def}{=} \frac{1}{n}\sum_{k=1}^{n}(x_k-\bar{x}_n)(y_k-\bar{y}_n)^{\intercal}. \tag{3.23}$$

where $\intercal$ *is the* transpose operator. *Note that in case* $N=1$ *the* cross-covariance *of the data sets* $\mathbf{x}$ *and* $\mathbf{y}$ *reduces to the real number*

$$s_{\mathbf{x},\mathbf{y},n} \stackrel{def}{=} \frac{1}{n}\sum_{k=1}^{n}(x_k-\bar{x}_n)(y_k-\bar{y}_n). \tag{3.24}$$

Definition 117 *Assume* $\det_M\left(\operatorname{diag}\left(s^2_{\mathbf{x},n}\right)\right)>0$ *and* $\det_N\left(\operatorname{diag}\left(s^2_{\mathbf{y},n}\right)\right)>0$, *where* $\operatorname{diag}\left(s^2_{\mathbf{x},n}\right)$ *[resp.* $\operatorname{diag}\left(s^2_{\mathbf{y},n}\right)$*] is the diagonal matrix obtainend by* $s^2_{\mathbf{x},n}$ *[resp.* $s^2_{\mathbf{y},n}$*] considering the corresponding diagonal entries and* $\det_M:\mathbb{R}^M\times\mathbb{R}^M\to\mathbb{R}$ *[resp.* $\det_N:\mathbb{R}^N\times\mathbb{R}^N\to\mathbb{R}$*] is the determinant function on* $\mathbb{R}^M\times\mathbb{R}^M$ *[resp.* $\mathbb{R}^N\times\mathbb{R}^N$*]. We call the* cross-correlation *of the data sets* $\mathbf{x}$ *and* $\mathbf{y}$ *the matrix in* $\mathbb{R}^{M\times N}$ *given by*

$$r_{\mathbf{x},\mathbf{y},n} \stackrel{def}{=} \operatorname{diag}\left(s^2_{\mathbf{x},n}\right)^{-1} s_{\mathbf{x},\mathbf{y},n} \operatorname{diag}\left(s^2_{\mathbf{y},n}\right)^{-1}. \tag{3.25}$$

Note that in case $M=N=1$ *the* cross-correlation *of the data sets* $\mathbf{x}$ *and* $\mathbf{y}$ *reduces to the real number*

$$r_{\mathbf{x},\mathbf{y},n} = \frac{s_{\mathbf{x},\mathbf{y},n}}{s_{\mathbf{x},n}s_{\mathbf{y},n}}. \tag{3.26}$$

Exercise 118 *With reference to Exercise* **??**, *compute the cross-covariance and cross-correlation of the height and weight data sets.*

Remark 119 *In case* $M=N=1$, *we have*

$$s_{\mathbf{x},\mathbf{y},n} = \frac{1}{n}\sum_{k=1}^{n}x_k y_k^{\intercal} - n\bar{x}_n\bar{y}_n$$

and

$$r_{\mathbf{x},\mathbf{y},n} = \frac{\sum_{k=1}^{n}x_k y_k - n\bar{x}_n\bar{y}_n}{\sqrt{\sum_{k=1}^{n}(x_k-\bar{x}_n)^2\sum_{k=1}^{n}(y_k-\bar{y}_n)^2}}.$$

Proposition 120 *In case* $M=N=1$, *we have*

$$-1 \le r_{\mathbf{x},\mathbf{y},n} \le 1. \tag{3.27}$$

Proof. We need to prove (3.27) only in the case $s_{\mathbf{x},n}s_{\mathbf{y},n}>0$. Now, by virtue of the Cauchy-Schwarz inequality (1.5), we can write

$$\left|\sum_{k=1}^{n}(x_k-\bar{x}_n)(y_k-\bar{y}_n)\right| \le \sum_{k=1}^{n}\left|(x_k-\bar{x}_n)(y_k-\bar{y}_n)\right| \le \left(\sum_{k=1}^{n}(x_k-\bar{x}_n)^2\right)^{1/2}\left(\sum_{k=1}^{n}(y_k-\bar{y}_n)^2\right)^{1/2}.$$

That is

$$n\left|s_{\mathbf{x},\mathbf{y},n}\right| \leq \sqrt{n}s_{\mathbf{x},n}\sqrt{n}s_{\mathbf{y},n}. \tag{3.28}$$

Dividing both the sides of (3.28) by $ns_{\mathbf{x},n}s_{\mathbf{y},n}$, we Then, obtain

$$\frac{\left|s_{\mathbf{x},\mathbf{y},n}\right|}{s_{\mathbf{x},n}s_{\mathbf{y},n}} \leq 1,$$

which is the desired (3.27). □

Definition 121 *When $r_{\mathbf{x},\mathbf{y},n} \neq 0$, [resp. $r_{\mathbf{x},\mathbf{y},n} = 0$] we say that the data sets $\mathbf{x}$ and $\mathbf{y}$ are* correlated *[resp.* uncorrelated*]. In case the data sets $\mathbf{x}$ and $\mathbf{y}$ are correlated, when $r_{\mathbf{x},\mathbf{y},n} > 0$ [resp. $r_{\mathbf{x},\mathbf{y},n} < 0$] we say that the data sets $\mathbf{x}$ and $\mathbf{y}$ are* positively *[resp.* negatively*] correlated. In particular, when $r_{\mathbf{x},\mathbf{y},n} = 1$ [resp. $r_{\mathbf{x},\mathbf{y},n} = -1$] we say that the data sets $\mathbf{x}$ and $\mathbf{y}$ are* perfectly positively *[resp.* perfectly negatively*] correlated.*

Remark 122 *In case $M = N = 1$, assume that we have*

$$y_k = \alpha x_k + \beta, \quad \forall k = 1, \ldots, n,$$

for $\alpha, \beta \in \mathbb{R}$, such that $\alpha \neq 0$. Then

$$r_{\mathbf{x},\mathbf{y},n} = \operatorname{sgn}(\alpha),$$

where sgn *is the signum operator.*

Proposition 123 *In case $M = N = 1$, let $(u_k)_{k=1}^n \equiv \mathbf{u}$ and $(v_k)_{k=1}^n \equiv \mathbf{v}$, be data sets of length n given by*

$$u_k \overset{def}{=} \alpha x_k + \beta \quad \text{and} \quad v_k \overset{def}{=} \gamma x_k + \delta, \quad \forall k = 1, \ldots, n,$$

for $\alpha, \beta, \gamma, \delta \in \mathbb{R}$, such that $\alpha, \gamma \neq 0$. Then, we have

$$s_{\mathbf{u},\mathbf{v},n} = \alpha\gamma s_{\mathbf{x},\mathbf{y},n}$$

and

$$r_{\mathbf{u},\mathbf{v},n} = \operatorname{sgn}(\alpha\gamma)\, r_{\mathbf{x},\mathbf{y},n},$$

where sgn *is the signum operator.*

3.6 Autocovariance and Autocorrelation

Let $(x_k)_{k=1}^n \equiv \mathbf{x}$ be an N-variate real data set of length n.

Definition 124 (Autocovariance of a data set) *We call the* autocovariance *of $\mathbf{x}$ at lag (shift) τ the matrix in $\mathbb{R}^N \times \mathbb{R}^N$ given by*

$$c_{\mathbf{x},n}(\tau) \overset{def}{=} \frac{1}{n}\sum_{k=1}^{n-\tau}(x_k - \bar{x}_n)(x_{k+\tau} - \bar{x}_n)^{\intercal}, \quad \forall \tau = 0, 1, \ldots, n-1. \tag{3.29}$$

In case $N = 1$, the autocovariance of $\mathbf{x}$ at lag (shift) τ reduces to the real number

$$c_{\mathbf{x},n}(\tau) \overset{def}{=} \frac{1}{n}\sum_{k=1}^{n-\tau}(x_k - \bar{x}_n)(x_{k+\tau} - \bar{x}_n), \quad \forall \tau = 0, 1, \ldots, n-1. \tag{3.30}$$

Remark 125 (Autocovariance of a data set) *We have*

$$c_{\mathbf{x},n}(0) = s^2_{\mathbf{x},n}, \tag{3.31}$$

where $s^2_{\mathbf{x},n}$ *is the variance-covariance matrix of* $\mathbf{x}$.

Definition 126 (Autocorrelation of a data set) *Fixed any* $\tau \in \mathbb{N}$, *assume that* $\det\left(\operatorname{diag}\left(s^2_{\mathbf{x},n}\right)\right) > 0$, *where* $\operatorname{diag}\left(s^2_{\mathbf{x},n}\right)$ *is the diagonal matrix having for diagonal entries the corresponding diagonal entries of* $s^2_{\mathbf{x},n}$ *and* $\det : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ *is the determinant function on* $\mathbb{R}^N \times \mathbb{R}^N$. *We call the* autocorrelation *of* $\mathbf{x}$ *at lag (shift)* τ *the matrix in* $\mathbb{R}^N \times \mathbb{R}^N$ *given by*

$$r_{\mathbf{x},n}(\tau) \overset{def}{=} \operatorname{diag}\left(s^2_{\mathbf{x},n}\right)^{-\frac{1}{2}} c_{\mathbf{x},n}(\tau) \operatorname{diag}\left(s^2_{\mathbf{x},n}\right)^{-\frac{1}{2}}, \quad \forall \tau = 0, 1, \ldots, n-1. \tag{3.32}$$

In case $N = 1$, *the autocorrelation of* $\mathbf{x}$ *at lag* τ *reduces to the real number*

$$r_{\mathbf{x},n}(\tau) = \frac{c_{\mathbf{x},n}(\tau)}{s^2_{\mathbf{x},n}}. \tag{3.33}$$

Remark 127 (Autocorrelation of a data set) *We have*

$$r_{\mathbf{x},n}(0) = r_{\mathbf{x},n}, \tag{3.34}$$

where $r_{\mathbf{x},n}$ *is the correlation matrix of* $\mathbf{x}$.

Proposition 128 (Autocorrelation of a real data set) *In case* $N = 1$, *we have*

$$|r_{\mathbf{x},n}(\tau)| \leq r_{\mathbf{x},n}(0) = 1, \tag{3.35}$$

for every $\tau = 1, \ldots, n-1$.

Proof. We have clearly

$$\sum_{k=1}^{n} (x_k - \bar{x}_n)^2 \geq \sum_{k=1}^{n-\tau} (x_k - \bar{x}_n)^2 \qquad \text{and} \qquad \sum_{k=1}^{n} (x_k - \bar{x}_n)^2 \geq \sum_{k=1}^{n-\tau} (x_{k+\tau} - \bar{x}_n)^2 .$$

It follows

$$2\sum_{k=1}^{n} (x_k - \bar{x}_n)^2 \geq \sum_{k=1}^{n-\tau} (x_k - \bar{x}_n)^2 + \sum_{k=1}^{n-\tau} (x_{k+\tau} - \bar{x}_n)^2 = \sum_{k=1}^{n-\tau} \left((x_k - \bar{x}_n)^2 + (x_{k+\tau} - \bar{x}_n)^2\right).$$

On the other hand,

$$(x_k - \bar{x}_n)^2 + (x_{k+\tau} - \bar{x}_n)^2 \geq 2\,|x_k - \bar{x}_n|\,|x_{k+\tau} - \bar{x}_n|$$

for every $\tau = 1, \ldots, n-1$ and every $k = 1, \ldots, n-\tau$. Therefore,

$$\sum_{k=1}^{n} (x_k - \bar{x}_n)^2 \geq \sum_{k=1}^{n-\tau} |x_k - \bar{x}_n|\,|x_{k+\tau} - \bar{x}_n| ,$$

for every $\tau = 1, \ldots, n-1$. As a consequence,

$$|c_{\mathbf{x},n}(\tau)| \leq \frac{1}{n}\sum_{k=1}^{n-\tau} |x_k - \bar{x}_n|\,|x_{k+\tau} - \bar{x}_n| \leq \frac{1}{n}\sum_{k=1}^{n} (x_k - \bar{x}_n)^2 = s^2_{\mathbf{x},n}.$$

for every $\tau = 1, \ldots, n-1$. The latter clearly implies

$$\frac{|c_{\mathbf{x},n}(\tau)|}{s^2_{\mathbf{x},n}} \leq 1,$$

for every $\tau = 1, \ldots, n-1$. On account of Definition 105, in case $N = 1$, and Equation (3.34), the desired result immediately follows. □

3.7 Frequency Tables, Graphs, Pie Charts

When a data set $(x_k)_{k=1}^n \equiv \mathbf{x}$ contains only a relatively small number of distinct values, it may be convenient to represent it by a *frequency* [rep. *relative frequency*] *table*. Data from a frequency [rep. relative frequency] table can be graphically pictured by a *line graph*, which plots the successive values from the set of values $\{x_j\}_{j=1}^n$ of the data set x on the horizontal axis and indicates the corresponding frequency f_j [rep. relative frequency r_j] (see Definition **??**) by the height of a vertical line.

A *pie chart* is often used to plot relative frequencies when the data are categorical that is non-numeric. A circle is drawn and sliced up into distinct sectors, each of them corresponding to a data value x_j, for every $j = 1, \ldots, m$. The area of the sector corresponding to the data value x_j is Then, as large as the relative frequency r_j of the data value.

Exercise 129 *With reference to Exercise* **??**, *write the frequency table and draw the graph of both the height and weight data sets.*

3.8 Steam-and-Leaf Plots

3.9 Boxplots and Outliers

A *boxplot* is a pictorial summary of a data set rather useful to describe the most preminent features of the data set. To introduce it, we need the following definition.

Let $(x_k)_{k=1}^n \equiv \mathbf{x}$ be a data set of length $n \geq 4$ and let $x_{1/2}$ be the median of x. Separate the data set in two halves x^- and x^+, which are the half of the data lower than the median and the the half of the data larger than the median, respectively.

$$\mathbf{x}^- \equiv \left\{x_k \in \mathbf{x} : x_k \leq x_{1/2}\right\} \quad \text{and} \quad \mathbf{x}^+ \equiv \left\{x_k \in \mathbf{x} : x_k \geq x_{1/2}\right\}.$$

Note that if n is odd [resp. even] the median $x_{1/2}$ belongs to both [resp. possibly none of] $\mathbf{x}^-$ and $\mathbf{x}^+$.

Definition 130 *We call* lower *[resp.* upper*]* fourth *and denote it by $x_{1/2}^-$ [resp. $x_{1/2}^+$] the median of $\mathbf{x}^-$ [resp. $\mathbf{x}^+$]. We call the* fourth spread *the real number f_s given by*

$$f_s \stackrel{def}{=} x_{1/2}^+ - x_{1/2}^-.$$

Note that the fourth spread f_s is not affected by the position of the data which are in the smallest or in the largest 25% of the data set.

Definition 131 *A point in the data set $\mathbf{x}$ which is farther than $1.5f_s$ from the closest fourth is called an* outlier*. An outlier is said to be* extreme *if it is farther than $3f_s$ from the closest fourth. It is said to be* mild *otherwise.*

Notation 132 *Write*

$$\hat{x} \equiv \min \mathbf{x} \quad \textit{and} \quad \check{x} \equiv \max \mathbf{x}.$$

A boxplot is based on the following measure summay

$$\hat{x}, \qquad x^{-}_{1/2}, \qquad x_{1/2}, \qquad x^{+}_{1/2}, \qquad \check{x}.$$

To draw a boxplot, draw a horizontal or vertical measurement scale and parallel to this draw a horizontal [resp. vertical] line which extends from the position $\hat{x}$ to $\check{x}$ on the measuremant scale. On the horizontal [resp. vertical] line, at the positions corresponding to $x^{-}_{1/2}$ and $x^{+}_{1/2}$ on the measurement scale, draw two small orthogonal segments of the same length to build a rectangular box of width f_s. In the box, at the position corresponding to $x_{1/2}$ on the measurement scale draw a further orthogonal segment to split the box in two parts. In the end, represent each mild [resp. extreme] outlier on the horizontal line by a full [empty] small circle at the position corresponding to the position of the outiler on the measurement scale.

The position of the median segment relative to the two edges of the box conveys information about the skewness of the 50% of the data set. The "whiskers" constituted by the parts of the horizontal line extending beyond the edges of the box to $\hat{x}$ and $\check{x}$, with the possible addition of full and empty small circles, convey information about the dispersion of the remaining 50% of the data and about the possible presence of outliers.

3.10 Histograms

For some data set $(x_k)_{k=1}^{n} \equiv \mathbf{x}$ the number of distinct values is too large or their (relative) frequency is too small to exploit a (relative) frequency table representation. In these cases, it is customary to group the data values along the horizontal axis in disjoint *class intervals* or *bins* and then, with reference to the vertical axis, plot a bar representing the number or the proportion of the data whose value falls in each bin. Such a bar graph plot is called a *histogram*. More specifically, we talk of *frequency* or *relative frequency* histogram according to whether we plot the number or the proportion of the data whose value falls in each bin. The endpoints of a bin are called *bin boundaries*. We will adopt the *left-end inclusion convention* which stipulates that each bin contains its left-end but not its right-end boundary; that is each bin is a closed-open interval. Depending on the actual data distribution and the goals of the analysis, different numbers of bins, or equivalently different bin widths, may be appropriate. The number of bins chosen should be a trade-off between choosing too few large bins, at the cost of losing too much information about the actual data values in a bin, and choosing too many small bins, which would result in a too small number or proportion of the data values in each bin to obtain a discernible pattern. In many cases, the appropriate number of bins may be found only through various attempts. However, there are various useful guidelines and rules of thumb (see e.g. [**?**, sec. 5.6]).

If the bins' width w is somehow assigned in advance Then, the number k of bins is determined by the formula

$$k = \left\lceil \frac{\max \mathbf{x} - \min \mathbf{x}}{w} \right\rceil ,$$

where $\max x$ [resp. $\min x$] is the maximum [resp. minimum] data value and $\lceil \cdot \rceil$ is the ceiling function.

If the bin's width is not assigned in advance a rather common choice of the number k of bins is determined by the *Tukey & Mosteller square-root rule* (1977)

$$k = \left\lceil \sqrt{n} \right\rceil ,$$

where n is the length of the data set . Another common choice of the number k of bins is determined by the *Sturges rule* (1926)

$$k = \lceil 1 + \log_2(n) \rceil,$$

where $\log_2(\cdot)$ is the base 2 logarithm. The Sturges rule is derived from a binomial distribution and implicitly assumes an approximately normal distribution. The *Teller & Scott rice rule* (1985)

$$k = \left\lceil \sqrt[3]{2n} \right\rceil,$$

is also commonly used as a simple alternative to the Sturges rule. The *Doane rule* (1976)

$$k = 1 + \log_2(n) + \log_2\left(1 + |skew_{\mathbf{x},n}| \sqrt{\frac{(n+1)(n+3)}{6(n-2)}}\right),$$

where $skew_{\mathbf{x},n}$ is the skewness of the data set $\mathbf{x}$, is a modification of the Sturges rule which attempts to improve the choice of the number k of bins with non-normal data. Another choice of the number k of bins is determined by the *Wichard rule* (2008)

$$k = 1 + \log_2(n) + \log_2\left(1 + kurt_{\mathbf{x},n}\sqrt{\frac{n}{6}}\right),$$

where $kurt_{\mathbf{x},n}$ is the kurtosis of the data set $\mathbf{x}$. In addition, there are rules aimed to the determination of the bin width w. The *Scott normal reference rule* (1979)

$$w = \frac{3.49 s_{\mathbf{x},n}}{\sqrt[3]{n}},$$

where $s_{\mathbf{x},n}$ is the standard deviation of the data set $\mathbf{x}$. The Scott normal reference rule is optimal for random samples of normally distributed data, in the sense that it minimizes the integrated mean squared error of the density estimate. The *Freedman & Diaconis rule* (1981)

$$w = \frac{2IQR}{\sqrt[3]{n}},$$

replacing the quantity 3.49σ of the Scott rule with the interquartile range IQR which is less sensitive than the standard deviation to outliers in data (see `https://en.wikipedia.org/wiki/Histogram`, see also `https://estatistics.eu/what-is-statistics-graph-figures-histogram/`).

Definition 133 *An histogram is said to be* approximately normal *if it has the following properties:*

1. the highest (relative) frequency of data is approximately attained at the middle bin;
2. the histogram is approximately symmetric about the middle bin;
3. moving from the middle bin in either direction, the height decreases in such a way that the entire histogram is approximately bell shaped.

Definition 134 *We say that a data set is* approximately normal *if it is possible to draw a histogram of the data set which is approximately normal.*

Remark 135 *If a data set is approximately normal, Then, the sample mean and the sample median of the data set are approximately equal.*

Empirical Rule Assume a data set is approximately normal with sample mean $\bar{x}_N$ and sample standard deviation s_N. Then, the following are true:

1. approximately 68% of the data values lie in the interval $[\bar{x}_n - s_{\mathbf{x},n}, \bar{x}_n + s_{\mathbf{x},n}]$;
2. approximately 95% of the data values lie in the interval $[\bar{x}_n - 2s_{\mathbf{x},n}, \bar{x}_n + 2s_{\mathbf{x},n}]$;
3. approximately 99.7% of the data values lie in the interval $[\bar{x}_n - 3s_{\mathbf{x},n}, \bar{x}_n + 3s_{\mathbf{x},n}]$.

Exercise 136 *With reference to Exercise* **??**, *determine the number and width of the beans according to some of the rules presented above and draw the corresponding histograms of both the height and weight, data sets. Are such histograms normal? May you guess the reason of their shape?*

3.11 Probability Plots

In data sets context, a P-P plot [resp. Q-Q plot], where PP stands for *probability probability* or *percent percent* [resp. QQ stands for *quantile quantile*], is a graphical method in the Cartesian plane $\mathbb{R}^2$ to compare the empirical probability distribution of a data set with the distribution of a random variable or another data set by plotting the respective distribution functions [the quantiles of the respective distributions] against each other.

3.11.1 Distribution Function of a Real Data Set

Let $(x_k)_{k=1}^n \equiv \mathbf{x}$ be a real data set of length n, let $\left(x_{(k)}\right)_{k=1}^n \equiv \mathbf{x}_{()}$ be the order statistic of $\mathbf{x}$, and let $\left(x_{\{j\}}\right)_{k=1}^m \equiv \mathbf{x}_{\{\}}$ the set of values of $\mathbf{x}$, indexed according to the natural order.

Definition 137 *We call the* empirical distribution function *of* $\mathbf{x}$ *the function* $F_{\mathbf{x}} : \mathbb{R} \to \mathbb{R}$ *given by*

$$F_{\mathbf{x}}(x) \overset{def}{=} \frac{1}{n} \sum_{k=1}^n 1_{\left[x_{(k)}, +\infty,\right)}(x), \quad \forall x \in \mathbb{R} \tag{3.36}$$

where $1_{\left[x_{(k)}, +\infty,\right)} : \mathbb{R} \to \mathbb{R}$ *is the indicator function of the half line* $\left[x_{(k)}, +\infty,\right)$, *for every* $k = 1, \ldots, n$.

Remark 138 *Setting*

$$\Omega_j \equiv \left\{k \in \{1, \ldots, n\} : x_k \leq x_{\{j\}}\right\}, \quad \forall j = 1, \ldots, m, \tag{3.37}$$

we have

$$F_{\mathbf{x}}(x) = \begin{cases} 0, & \text{if } x < x_{\{1\}}, \\ \frac{|\Omega_j|}{n}, & \text{if } x_{\{j\}} \leq x < x_{\{j+1\}}, \\ 1, & \text{if } x_{\{m\}} \leq x, \end{cases} \tag{3.38}$$

where $|\Omega_j|$ *is the cardinality of the set* Ω_j.

3.11.2 Density Function of a Real Data Set

Let $(x_k)_{k=1}^n \equiv \mathbf{x}$ be a real data set of length n, let $\left(x_{(k)}\right)_{k=1}^n \equiv \mathbf{x}_{()}$ be the order statistic of $\mathbf{x}$, and let $(B_j)_{j=1}^m \equiv \mathcal{B}$ a partition of the interval $\left[x_{(1)}, x_{(n)}\right]$.

Definition 139 *We call the* empirical density function *of* $\mathbf{x}$*, given the partition* $\mathcal{B}$*, the function* $f_{\mathbf{x}} : \mathbb{R} \to \mathbb{R}$ *given by*

$$f_{\mathbf{x}}(x) \stackrel{def}{=} \frac{1}{n} \sum_{j=1}^m \frac{\sum_{k=1}^n 1_{B_j}\left(x_{(k)}\right)}{\mu_L(B_j)} 1_{B_j}(x), \quad \forall x \in \mathbb{R}. \tag{3.39}$$

3.11.3 P-P plots

Let $F_X : \mathbb{R} \to \mathbb{R}$ be the distribution function of a real random variable X and let $(y_k)_{k=1}^n \equiv \mathbf{y}$ be a data set of length n with order statistic $\left(y_{(k)}\right)_{k=1}^n \equiv \mathbf{y}_{()}$ and empirical distribution function $F_{\mathbf{y}} : \mathbb{R} \to \mathbb{R}$.

Definition 140 *We call the distribution function* F_X *[resp. the data set* $\mathbf{y}$*] the* reference *distribution [resp. the* test *data set].*

Definition 141 *We call a* P-P plot *of* $\mathbf{y}$ *against* X*, the representation in the Cartesian plane* $\mathbb{R}^2$ *of the parametric curve* $PP_{X,\mathbf{y}} : \{1, \ldots, n\} \to \mathbb{R}^2$ *given by*

$$PP_{X,\mathbf{y}}(k) \stackrel{def}{=} \left(F_X\left(y_{(k)}\right), F_{\mathbf{y}}\left(y_{(k)}\right)\right), \quad \forall k \in \{1, \ldots, n\}. \tag{3.40}$$

Remark 142 *Assume the test data set* $\mathbf{y}$ *is drawn from the same probability distribution of* X*. Then, the pattern of the corresponding P-P plot* $PP_{X,y}$ *is very close to the straight line* $\mathbf{y} = x$*. Conversely, assume that the pattern of the P-P plot* $PP_{X,\mathbf{y}}$ *is very close to the straigth line* $y = x$*. Then, the test data set* $\mathbf{y}$ *is likely drawn from the same probability distribution of* X*.*

Remark 143 *Assume the test data set* $\mathbf{y}$ *is drawn from a probability distribution which is more concentrated [resp. dispersed] than the distribution of* X*. Then, the pattern of the corresponding P-P plot* $PP_{X,\mathbf{y}}$ *is steeper [resp. flatter] than the straigth line* $y = x$*. Conversely, assume that the pattern of the P-P plot* $PP_{X,\mathbf{y}}$ *is steeper [resp. flatter] than the straigth line* $y = x$*. Then, the test data set* $\mathbf{y}$ *is likely drawn from a probability distribution which is more concentrated [resp. dispersed] than the distribution of* X*.*

Remark 144 *Assume the test data set* $\mathbf{y}$ *is drawn from a probability distribution which is more concentrated on the left [resp. right] than the distribution of* X*. Then, the pattern of the corresponding P-P plot* $PP_{X,\mathbf{y}}$ *is arched downwards [resp. upwards]. Conversely, assume that the pattern of the PP plot* $PP_{X,\mathbf{y}}$ *arched downwards [resp. upwards]. Then, the test data set* $\mathbf{y}$ *is likely drawn from a probability distribution which is more concentrated on the left [resp. right] than the distribution of* X*.*

Remark 145 *Assume the test data set* $\mathbf{y}$ *is drawn from a probability distribution which has lighter [resp. heavier] tails than the distribution of* X*. Then, the pattern of the corresponding P-P plot* $PP_{X,\mathbf{y}}$ *is "S" [resp. reverse "S"]-shaped like the logistic [resp. logit]. Conversely, assume that the pattern of the P-P plot* $PP_{X,\mathbf{y}}$ *is "S" [resp. reverse "S"]-shaped like the logistic [resp. logit]. Then, the test data set* $\mathbf{y}$ *is likely drawn from a distribution which has lighter [resp. heavier] tails than the distribution of* X*.*

3.11.4 Q-Q plots

Given a real random variable X with strictly increasing and continuous distribution function $F_X : \mathbb{R} \to \mathbb{R}$, recall that the quantile of order q or q-quantile of the probability distribution of X is the real number x_q which fulfills the equation

$$F_X(x_q) = q,$$

on varying of $q \in (0,1)$ (see Definitions 523 and Proposition **??**). Such a q-quantile is also called the $100q$th percentile of the distribution X. Given a real data set $(x_k)_{k=1}^n \equiv \mathbf{x}$ of length n and writing $\left(x_{(k)}\right)_{k=1}^n \equiv \mathbf{x}_{()}$ for the order statistic of $\mathbf{x}$, recall that the quantile of order q or *q-quantile* of $\mathbf{x}$, also called the $100q$th *percentile*, is the real number x_q given by

$$x_q \overset{\text{def}}{=} x_{(\lfloor h \rfloor)} + (h - \lfloor h \rfloor)\left(x_{(\lfloor h \rfloor + 1)} - x_{(\lfloor h \rfloor)}\right),$$

where $h \equiv (n+1)\,p$, on varying of $p \in (0,1) \cap \mathbb{Q}$ (see Definition 89). However, in a computational context, the $(k - 0.5)/n$*-quantile* or $100\,(k - 0.5)/n$ th *percentile* of the data set $\mathbf{x}$ is intended to be the real number $x_{(k)}$ which occupies the kth place in the order statistics $\mathbf{x}_{()}$ of $\mathbf{x}$, on varying of $k = 1, \ldots, n$ (see Definition 95).

Ler X, Y real random variables with strictly increasing and continuous distribution function $F_X : \mathbb{R} \to \mathbb{R}$ and $F_Y : \mathbb{R} \to \mathbb{R}$, respectively.

Definition 146 *We call the* Q-Q plot of Y against X*, the representation in the Cartesian plane* $\mathbb{R}^2$ *of the parametric curve* $QQ_{X,Y} : (0,1) \to \mathbb{R}^2$ *given by*

$$QQ_{X,Y} \overset{def}{=} (x_q, y_q) \quad \forall q \in (0,1),$$

where x_q *and* y_q *fulfill the equations*

$$F_X(x_q) = F_Y(y_q) = q.$$

Let $(x_k)_{k=1}^n \equiv x$ and $(y_k)_{k=1}^n \equiv y$ two data set of the same length n and let $\left(x_{(k)}\right)_{k=1}^n \equiv \mathbf{x}_{()}$ and $\left(y_{(k)}\right)_{k=1}^n \equiv \mathbf{y}_{()}$ be the order statistics of $\mathbf{x}$ and $\mathbf{y}$, respectively. We agree that $\mathbf{x}$ [resp. $\mathbf{y}$] is a set of data generated from a reference [resp. tested] probability distribution.

Definition 147 *We call the data set* $\mathbf{x}$ *[resp.* $\mathbf{y}$*] the* reference *[resp.* test*] data set.*

Definition 148 *We call the* Q-Q plot of $\mathbf{y}$ against $\mathbf{x}$*, the representation in the Cartesian plane* $\mathbb{R}^2$ *of the parametric point curve* $QQ_{x,y} : \{1, \ldots, n\} \to \mathbb{R}^2$ *given by*

$$QQ_{\mathbf{x},\mathbf{y}}(k) \overset{def}{=} \left(x_{(k)}, y_{(k)}\right) \quad \forall k \in \{1, \ldots, n\}.$$

Analogously to the case of a P-P plot, the shape of the pattern of a Q-Q plot provides us with several information about the probability distributions which generates the reference and test data.

Remark 149 *Assume the test data set* $\mathbf{y}$ *is drawn from the same probability distribution generating the reference data set* $\mathbf{x}$*. Then, the pattern of the corresponding Q-Q plot* $QQ_{\mathbf{x},\mathbf{y}}$ *is very close to the straigth line* $y = x$*. Conversely, assume that the pattern of the Q-Q plot* $QQ_{\mathbf{x},\mathbf{y}}$ *is very close to the straigth line* $y = x$*, Then, the test data set* $\mathbf{y}$ *is likely drawn from the same probability distribution which generates the reference data set* $\mathbf{x}$*.*

Remark 150 *Assume the test data set* $\mathbf{y}$ *is drawn from a probability distribution which is a linear transformation of the distribution generating the reference data set* $\mathbf{x}$*. Then, the pattern of the corresponding Q-Q plot* $QQ_{\mathbf{x},\mathbf{y}}$ *is very close to a straight line. Conversely, assume that the pattern of the Q-Q plot* $QQ_{\mathbf{x},\mathbf{y}}$ *is very close to a straight line, Then, the test data set* $\mathbf{y}$ *is likely drawn from a linear transformation of the probability distribution which generates the reference data set* $\mathbf{x}$*.*

Remark 151 *Assume the data sets* $\mathbf{x}$ *and* $\mathbf{y}$ *are drawn from the same family of probability distributions but it is not true that the distribution which generates the test data set* $\mathbf{y}$ *is a linear transformation of the distribution generating the reference data set* $\mathbf{x}$*. Then, the differences in the characterizing parameters of the probability distributions which generate* $\mathbf{x}$ *and* $\mathbf{y}$ *do not generally allow a straight line shape of the pattern of the Q-Q plot* $QQ_{\mathbf{x},\mathbf{y}}$*.*

Remark 152 *Assume the test data set* $\mathbf{y}$ *is drawn from a probability distribution which is more concentrated [resp. dispersed] than the distribution generating the reference data set* $\mathbf{x}$*. Then, the pattern of the corresponding Q-Q plot* $QQ_{\mathbf{x},\mathbf{y}}$ *is flatter [resp. steeper] than the straigth line* $y = x$*. Conversely, assume that the pattern of the Q-Q plot* $QQ_{\mathbf{x},\mathbf{y}}$ *is flatter [resp. steeper] than the straigth line* $y = x$*. Then, the test data set* $\mathbf{y}$ *is likely drawn from a probability distribution which is more concentrated [resp. dispersed] than the distribution generating the reference data set* $\mathbf{x}$*.*

Remark 153 *Assume the test data set* $\mathbf{y}$ *is drawn from a probability distribution which is more concentrated on the left [resp. right] than the generating the reference data set* $\mathbf{x}$*. Then, the pattern of the corresponding QQ plot* $QQ_{\mathbf{x},\mathbf{y}}$ *is arched downwards [resp. upwards]. Conversely, assume that the pattern of the Q-Q plot* $QQ_{\mathbf{x},\mathbf{y}}$ *is arched downwards [resp. upwards]. Then, the test data set* $\mathbf{y}$ *is likely drawn from a probability distribution which is more concentrated on the left [resp. right] than the distribution generating the reference data set* $\mathbf{x}$*.*

Remark 154 *Assume the test data set* $\mathbf{y}$ *is drawn from a probability distribution which has lighter [resp. heavier] tails than the generating the reference data set* x*. Then, the pattern of the corresponding Q-Q plot* $QQ_{\mathbf{x},\mathbf{y}}$ *is "S" [resp. reverse "S"]-shaped like the logistic [resp. logit]. Conversely, assume that the pattern of the Q-Q plot* $QQ_{\mathbf{x},\mathbf{y}}$ *is "S" [resp. reverse "S"]-shaped like the logistic [resp. logit]. Then, the test data set* $\mathbf{y}$ *is likely drawn from a probability distribution which has lighter [resp. heavier] tails the distribution generating the reference data set* $\mathbf{x}$*.*

Note that exchanging the test and the reference data sets results in a reversion of the shape of the associated Q-Q plot. In general a Q-Q plot is more sensitive to deviances from normality in the tails of the data set than the corresponding P-P plot, whereas a P-P plot is more sensitive to deviances near the mean of the distribution than the corresponding Q-Q plot. This makes a Q-Q plot a better detector of outliers in the test data set that a P-P plot. The identification of outliers in a data set is an important goal. Therefore, Q-Q plots are more frequently used than P-P plots to assess the normality of a data set.

3.12 Density Kernel Estimation

Let $(x_k)_{k=1}^n \equiv \mathbf{x}$ be an N-variate real data set of length n. Assume we know that $\mathbf{x}$ is generated by the realizations of an absolutely continuous random variable X. The goal of the technique known as *density kernel estimation* (*KDE*), after Emanuel Parzen and Murray Rosemblatt, is to infer the shape of the density $f_X : \mathbb{R} \to \mathbb{R}$ of X from the data set $\mathbf{x}$.

Definition 155 *We call* density kernel estimation *of* $f_{\mathbf{x}} : \mathbb{R} \to \mathbb{R}$ *with* bandwidht $h > 0$*, via the data set* $\mathbf{x}$*, the function* $\hat{f}_{\mathbf{x},h} : \mathbb{R} \to \mathbb{R}$ *given by*

$$\hat{f}_{\mathbf{x},h}(x) \overset{def}{=} \frac{1}{nh} \sum_{k=1}^{n} K\left(\frac{x_k - x}{h}\right), \quad \forall x \in \mathbb{R}, \tag{3.41}$$

where $K : \mathbb{R} \to \mathbb{R}$*, referred to as the* kernel*, is a given probability density which is symmetric about* 0 *and has finite second order moment. That is*

1. $\int_{\mathbb{R}} K(x)\, d\mu_L(x) = 1$*;*
2. $K(-x) = K(x)$*, for every* $x \in \mathbb{R}$*;*
3. $\int_{\mathbb{R}} x^2 K(x)\, d\mu_L(x) < \infty$*.*

Remark 156 *We have*

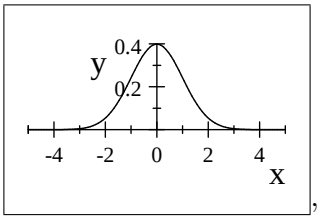
$$\sigma_K^2 = \int_{\mathbb{R}} x^2 K(x)\, d\mu_L(x).$$

Definition 157 *We call* efficiency *of the kernel the positive number*

$$EK \overset{def}{=} \left(\int_{\mathbb{R}} x^2 K(x)\, d\mu_L(x)\right)^{1/2} \int_{\mathbb{R}} K^2(x)\, d\mu_L(x). \tag{3.42}$$

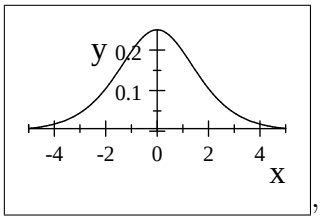
Among the kernels which are commonly used in *KDE* we recall

the *Gaussian kernel*: $K(x) \overset{\text{def}}{=} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}, \quad \forall x \in \mathbb{R}$



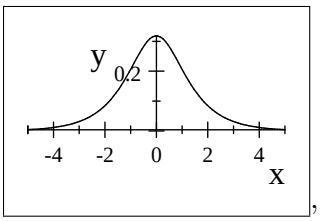
, $EK = 95.1\%$,

the *logistic kernel*: $K(x) \overset{\text{def}}{=} \frac{1}{2+e^x+e^{-x}}, \quad \forall x \in \mathbb{R},$



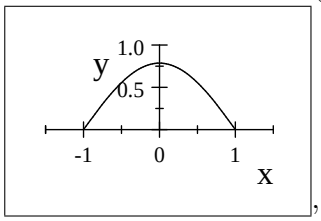
, $EK = 88.7\%$,

the *sigmoid kernel*: $K(x) \overset{\text{def}}{=} \frac{2}{\pi} \frac{1}{e^x+e^{-x}}, \quad \forall x \in \mathbb{R},$



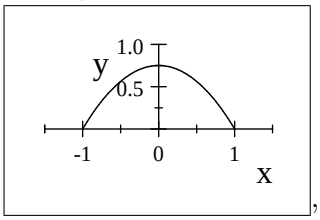
, $EK = 84.3\%$,

the *cosine kernel*: $K(x) \stackrel{\text{def}}{=} \frac{\pi}{4}\cos\left(\frac{\pi}{2}x\right)1_{[-1,1]}(x), \quad \forall x \in \mathbb{R},$



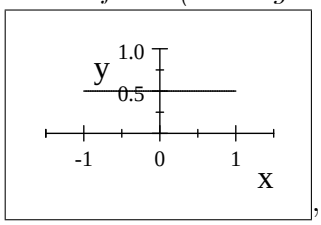
, $EK = 99.9\%$

the *Epanechnikov kernel*: $K(x) \stackrel{\text{def}}{=} \frac{3}{4}(1-x^2)1_{[-1,1]}(x), \quad \forall x \in \mathbb{R},$



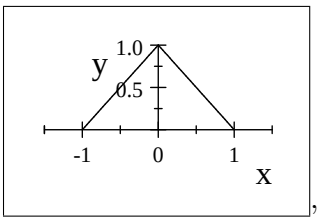
, $EK = 100\%,$

the *uniform (rectangular) kernel*: $K(x) \stackrel{\text{def}}{=} \frac{1}{2}1_{[-1,1]}(x), \quad \forall x \in \mathbb{R},$



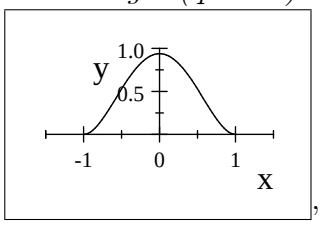
, $EK = 92,9\%,$

the *triangular kernel*: $K(x) \stackrel{\text{def}}{=} (1-|x|)1_{[-1,1]}(x), \quad \forall x \in \mathbb{R},$



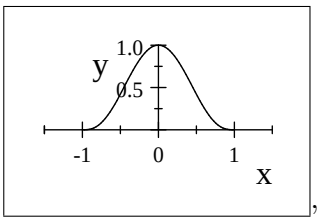
, $EK = 98,6\%,$

the *biweight (quartic) kernel*: $K(x) \stackrel{\text{def}}{=} \frac{15}{16}(1-x^2)^2 1_{[-1,1]}(x), \quad \forall x \in \mathbb{R},$



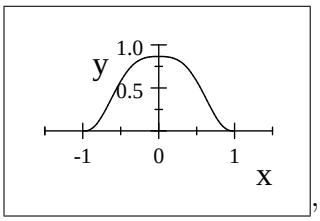
, $EK = 99.4\%,$

the *triweight kernel*: $K(x) \stackrel{\text{def}}{=} \frac{35}{35}(1-x^2)^3 1_{[-1,1]}(x), \quad \forall x \in \mathbb{R},$



, $EK = 98.7\%,$

the *tricube kernel*: $K(x) \stackrel{\text{def}}{=} \frac{70}{81}\left(1-|x|^3\right)^3 1_{[-1,1]}(x), \quad \forall x \in \mathbb{R},$



, $EK = 99.8\%$.

The bandwight h of a kernel is a free smoothing parameter which exerts a strong influence on the resulting estimate.

Theorem 158 *Assume the unknown density $f_{\mathbf{x}} : \mathbb{R} \to \mathbb{R}$ is twice differentiable with square integrable second derivative $f''_{\mathbf{x}} : \mathbb{R} \to \mathbb{R}$. Then, an optimal choice of the parameter h comes from the minimization of the following* asymptotic mean integrated squared error *(*AMISE*) function*

$$AMISE(h) \stackrel{def}{=} \frac{1}{nh}\int_{\mathbb{R}} K^2(x)\, d\mu_L(x) + \frac{1}{4}\sigma_k^2 h^4 \int_{\mathbb{R}} f''_{\mathbf{x}}(x)^2\, d\mu_L(x).$$

Remark 159 *The minimum of $AMISE(h)$ is achieved at*

$$\hat{h} = \left(\frac{\int_{\mathbb{R}} K(x)^2\, d\mu_L(x)}{n\sigma_K^2 \int_{\mathbb{R}} f''(x)^2\, d\mu_L(x)}\right)^{1/5}.$$

It follows

$$AMISE\left(\hat{h}\right) = O\left(n^{-4/5}\right).$$

However, the above determination of h cannot be used directly because it involves the second derivative of the unknown density function. Therefore a variety of data-based methods has been developed to select the bandwidth. If the random variable X which generates the data set $\mathbf{x}$ is approximatively Gaussian distributed and a Gaussian kernel is used to estimate the density of X, Then, the optimal choice of h is given by the following Silverman's rule of thumb.

$$h = \left(\frac{4}{3}\right)^5 \frac{s_{\mathbf{x},n}}{n^5},$$

were $s_{\mathbf{x},n}$ is the standard deviation of the data set $\mathbf{x}$. While the Silverman's rule of thumb is easy to compute, it should be use with caution as it can lead to inaccurate estimates when the density of $\mathbf{x}$ is not close to be normal.

Part III

Elements of Probability Theory

Chapter 4

Probability Spaces

A *random phenomenon* or *experiment* is any natural or social phenomenon or controlled experiment whose outcome cannot be forecasted with certainty by an observer, in light of the available information. Typical random phenomena are metereological or hydrologic phenomena, earthquakes, sunspots, opinion polls, fluctuations of asset prices in a financial maket, and so on. Typical random experiment are the flip of a coin, which depicts the most basic randomness, or the roll of a die, or the draw of a ball from an urn, or the pick of a card out of a deck, or the sampling from a population, and so on. In many cases, the randomness of a phenomenon is related to the observer's information: some phenomenon or experiment which is studied conveniently by applying a random model could also be studied by means of a deterministic model, provided all relevant information was available. On the other hand, the randomness of the phenomena concerning Quantum Physics seems to be a matter of fact. Eventually, as a consequence of the De Broglie wave-particle duality and the Heisenberg Uncertainty Principle, there is no information in light of which the outcome of a quantum experiment can be uniquely forecasted by any observer. Thus, chance seems to play a basic role on the stage of nature as well as space and time. However, an investigation of the role of chance in natural phenomena is far beyond the goal of these notes.

4.1 Sample Spaces

4.1.1 Outcomes and Events

By a sample space of a random phenomenon or experiment we mean the set of all possible *outcomes* of the random phenomenon or experiment whose occurrence or nonoccurrence can be established unambiguously by an observer. Using the standard notation, we write Ω for a sample space and ω for a generic outcome. We also refer to ω as a *sample point* of Ω.

A "gathering together into a whole" (see Appendix 19) of the outcomes of a random phenomenon or experiment is called an *event*. Hence, any event is naturally identified with a subset of the sample space Ω. In particular, the empty subset $\varnothing$ of Ω is an event, referred to as the *impossible event*, and the sample space Ω itself is an event, referred to as the *sure event*. Note that the impossible event is to represent the gathering together of contradictory outcomes, while the sure event is to represent the gathering together of all possible outcomes. Other particular events are the events made up by a single outcome $\omega \in \Omega$ represented by the subsets $\{\omega\}$ of Ω, on varying of $\omega \in \Omega$, and referred to as the *elementary events*, or the *singletons* of Ω. The family of all events is naturally identified with the family $\mathcal{P}(\Omega)$ of all subsets of Ω. The latter is

referred to as the *set of the parts* or *power set* of Ω and also denoted by 2^{Ω}. Given any event E we can equivalently write $E \subseteq \Omega$ or $E \in \mathcal{P}(\Omega)$. The identification between events of a random phenomenon or experiment and subsets of the corresponding sample space is pretty strict. By virtue of this identification, the logic rules for the combination of events are reflected in the rules for the manipulation of sets.

Let Ω be a sample space and let E and F be events of Ω.

Definition 160 *We say that E* (logically) implies *F, or F is* (logical) consequence *of E, if $E \subseteq F$ or, which is the same, $F \supseteq E$.*

Definition 161 *We say that E is* (logically) equivalent *to F, if $E = F$.*

Axiom 162 (Extensionality Principle) *The events E and F are logically equivalent if and only if*

$$E \subseteq F \quad \textit{and} \quad F \subseteq E. \tag{4.1}$$

Example 163 *We flip a coin. The only possible outcomes of the flip are the symbols "heads" and "tails". Let us denote them by 1 and 0, respectively. Thus, the sample space Ω can be represented by the set $\{1, 0\}$.*

Example 164 *We roll a die. Each of the six faces of the die is marked with spots: from one to six spots. Clearly, the possible outcomes of the roll can be represented by the numbers $1, \ldots, 6$ of the spots marking the faces. Thus, the sample space Ω can be represented by the set $\{1, \ldots, 6\}$. As a consequence, the event "the outcome of the roll is odd [resp. even]" is represented by the subset $E_{\mathbb{O}} \equiv \{1, 3, 5\}$ [resp. $E_{\mathbb{E}} \equiv \{2, 4, 6\}$] of Ω and the event "the outcome of the roll is j" is represented by the singleton $\{j\}$, for $j = 1, \ldots, 6$. Note that $\{1\} \subseteq E_{\mathbb{O}}$, but $E_{\mathbb{O}} \nsubseteq \{1\}$. Actually, it is true that "if the outcome of the roll is 1, Then, the outcome is odd", but it is false that "if the outcome of the roll is odd, Then, the outcome is 1". Otherwise saying, the event $\{1\}$ implies the event $E_{\mathbb{O}}$ (the event $E_{\mathbb{O}}$ is consequence of the event $\{1\}$), but the converse is not true. Similarly, $\{1\} \nsubseteq E_{\mathbb{E}}$, since it is false that "if the outcome of the roll is 1, Then, the outcome is even".*

Let Ω be a sample space and let E an event of Ω.

Definition 165 *We call the* complement *of E, the event E^c that occurs whenever the event E does not. In symbols,*

$$E^c \stackrel{def}{=} \{\omega \in \Omega : \omega \notin E\}.$$

Remark 166 *We have*

$$\Omega^c = \varnothing \quad \textit{and} \quad \varnothing^c = \Omega.$$

Let Ω be a sample space and let E and F be events of Ω.

Proposition 167 *We have*

1. $(E^c)^c = E$;
2. $E \subseteq F \Leftrightarrow F^c \subseteq E^c$.

Definition 168 *We call the* union *of E and F, the event $E \cup F$ that occurs when at least one of the events E and F occurs. In symbols,*

$$E \cup F \stackrel{def}{=} \{\omega \in \Omega : \omega \in E \ \vee \ \omega \in F\}.$$

Proposition 169 *We have*

1. $E \cup E = E$;
2. $(E \cup F) \cup G = E \cup (F \cup G)$;
3. $E \cup F = F \cup E$.

Proposition 170 *We have*

$$E \cup F = F \Leftrightarrow E \subseteq F.$$

In particular,

$$E \cup \Omega = E \Leftrightarrow E = \Omega \quad \textit{and} \quad E \cup \varnothing = \varnothing \Leftrightarrow E = \varnothing.$$

Definition 171 *We call the* intersection *of E and F, the event $E \cap F$ that occurs when both the events E and F occurs. In symbols,*

$$E \cap F \stackrel{def}{=} \{\omega \in \Omega : \omega \in E \ \wedge \ \omega \in F\}.$$

Proposition 172 *We have*

1. $E \cap E = E$;
2. $(E \cap F) \cap G = E \cap (F \cap G)$;
3. $E \cap F = F \cap E$.

Proposition 173 *We have*

$$E \cap F = E \Leftrightarrow E \subseteq F.$$

In particular,

$$E \cap \Omega = \Omega \Leftrightarrow E = \Omega \quad \textit{and} \quad E \cap \varnothing = E \Leftrightarrow E = \varnothing.$$

Definition 174 *We say that E and F are* jointly exhaustive *or* complete *if*

$$E \cup F = \Omega.$$

Definition 175 *We say that E and F are* incompatible *if*

$$E \cap F = \varnothing.$$

Definition 176 *We say that E and F constitute a* partition *of Ω if they are both jointly exhaustive and incompatible.*

The logical operations of complementation, union, and intersection and are connected by the *De Morgan* and *distributive* laws.

Proposition 177 (De Morgan laws) *We have*

1. $(E \cup F)^c = E^c \cap F^c$;
2. $(E \cap F)^c = E^c \cup F^c$.

Proposition 178 (Distributive laws) *Let G be another event of Ω. We have*

1. $E \cup (F \cap G) = (E \cup F) \cap (E \cup G)$;
2. $E \cap (F \cup G) = (E \cap F) \cup (E \cap G)$.

Definition 179 *We call the* difference *of F and E the event $F - E$ which occurs whenever F occurs and E does not. In symbols,*

$$F - E \overset{def}{=} \{\omega \in \Omega : \omega \in F \ \wedge \ \omega \notin E\}.$$

The difference $F - E$ is also called the relative complement *of E in F and denoted by the symbol E_F^c.*

Proposition 180 *We have*

$$F - E = F \cap E^c.$$

In particular,

$$\Omega - E = E^c.$$

Proposition 181 *We have*

$$F - E = F \Leftrightarrow F \cap E = \varnothing$$

and

$$F - E = \varnothing \Leftrightarrow F \subseteq E.$$

Definition 182 *We call* symmetric difference *of E and F, the event $E \Delta F$ that occurs whenever E occurs and F does not, or F occurs and E does not. In symbols,*

$$E \Delta F \overset{def}{=} \{\omega \in \Omega \mid (\omega \in E \ \wedge \ \omega \notin F) \vee (\omega \in F \ \wedge \ \omega \notin E)\}.$$

Proposition 183 *We have*

$$E \Delta F = (E - F) \cup (F - E),$$

where $E - F$ and $F - E$ are incompatible.

Example 184 *We flip a coin (see Example 163). The elementary events are the singletons $\{h\} \equiv H$ and $\{t\} \equiv T$. The event $H \cup T$, which means "the outcome of the flip is heads or tails", is the sure event Ω, and the event $H \cap T$, which means "the outcome of the flip is heads and tails", is the impossible event $\varnothing$. The family $\mathcal{P}(\Omega)$ of all events of Ω is clearly $\{\varnothing, \Omega, H, T\}$. The events H and T are a partition of Ω.*

Example 185 *We roll a die (see Example 164). The elementary events are the singletons* $\{j\}$ *for* $j = 1, \dots, 6$*. The event* $E_{\mathbb{O}} \cup E_{\mathbb{E}}$*, which means "the outcome of the throw is even or odd", is the sure event* Ω*. The event* $E_{\mathbb{O}} \cap E_{\mathbb{E}}$*, which means "the outcome of the throw is even and odd", is the impossible event* $\varnothing$*. Note that* $E_{\mathbb{O}}^c = E_{\mathbb{E}}$ *and the events* $E_{\mathbb{O}}$ *and* $E_{\mathbb{E}}$ *are a partition of* Ω*. Clearly, many other events are possible. For instance,* $\{1, 2, 3\}$*, which means "the outcome of the throw is not greater than* 3*", or* $\{4, 5, 6\}$*, which means "the outcome of the throw is not smaller than* 4*. By combinatorics, it is not difficult to prove that the family* $\mathcal{P}(\Omega)$ *of all events of* Ω *contains* $2^6 = 64$ *distinct events (see also 1 of Theorem 204).*

Example 186 *At a given instant* t *and for a given time interval* $\Delta t \geq 0$*, we check the number of people waiting in a queue, or the number of phone calls incoming to a telephone exchange, or the number of decaying atoms in a radioactive mass or the number of messages arriving at your favorite group chat. The outcome of the check is clearly a positive integer. To study this type of random phenomena it turns out to be convenient to choose the set of all natural numbers* $\mathbb{N} \equiv \{0, 1, 2, \dots\}$ *as the sample space* Ω*. The elementary events are* $E_n \equiv \{n\}$*, for* $n \in \mathbb{N}$*, each of which means "the outcome of the check is* n*". Other possible events are* $E_{\mathbb{O}} \equiv \{1, 3, 5, \dots\}$*, which means "the outcome of the check is an odd number", or* $E_{\mathbb{E}} \equiv \{0, 2, 4, \dots\}$*, which means "the outcome of the check is an even number". Here, in principle, very many events are possible. Eventually, the family* $\mathcal{P}(\Omega)$ *is uncountable (see 2 of Theorem 204).*

From Example 186, it is possible to realize that, in general, there is some arbitrariness in the choice of the sample space Ω which is invoked to represent the set of all possible outcomes of a random phenomenon or experiment. For instance, nobody would expect to count more than 10^5 people waiting in a bus queue in thirty minutes! But it is not impossible that 10^5 phone calls income to a large telephone exchange in thirty minutes, or that 10^7 atoms of a radioactive mass decay. However, as we will see, rather than modeling a specific random phenomenon by a specific sample space, the strategy is to use a general sample space and to assign a specific evaluation of the possibility of occurrence of its events, to say different *probability measures*, which tailor the general sample space to the specific random phenomenon.

Example 187 *At a given instant* t*, we check the market value of a share of stock, or the parity between two currencies. In this case, we should expect any positive rational number as the result of our check. Thus, the natural sample space should be the set* $\mathbb{Q}_+$ *of all positive rational numbers. However, for technical reasons that we will present later, it turns out more convenient to choose the set* $\mathbb{R}_+$ *of all positive real numbers as the sample space* Ω*. Moreover, if we agree to use a logarithmic scale to represent the outcomes of the check, Then, we are led to choice the set* $\mathbb{R}$ *of all real numbers as* Ω*. In this case, the family* $\mathcal{P}(\Omega) \equiv 2^{\Omega}$ *contains* $2^{\aleph_1} > \aleph_1$ *distinct events, and it turns out to be unmanageable for setting up a non-trivial probability measure. We will show how to overcome this difficulty.*

The definition of the logical operations of union and intersection can be easily extended to families of events of a sample space.

Let Ω be a sample space and let $\{E_j\}_{j \in J}$ be a family of events of Ω.

Definition 188 *We call the* union *of* $\{E_j\}_{j \in J}$*, the event* $\bigcup_{j \in J} E_j$ *that occurs when at least one of the events of the family* $\{E_j\}_{j \in J}$ *occurs. In symbols,*

$$\bigcup_{j \in J} E_j \stackrel{def}{=} \{\omega \in \Omega : \exists j \in J \text{ s.t. } \omega \in E_j\}.$$

Proposition 189 *We have*

1. $\bigcup_{j\in J} E_j = E$ *for any family* $\{E_j\}_{j\in J}$ *of events of* Ω *such that* $E_j = E$*, for every* $j \in J$*;*
2. $\left(\bigcup_{j\in J_1} E_j\right) \cup \left(\bigcup_{j\in J_2} E_j\right) = \bigcup_{j\in J} E_j$*, for all* $J_1, J_2 \subseteq J$ *such that* $J_1 \cup J_2 = J$*;*
3. $\bigcup_{j\in J} E_j = \bigcup_{j\in J} E_{\pi(j)}$ *for any family* $\{E_j\}_{j\in J}$ *of events of* Ω *and any bijection* $\pi : J \to J$*.*

Proof. . □

Proposition 190 *We have*

$$\bigcup_{j\in J} E_j = E_{j_0},\ \text{for some } j_0 \in J,\ \Leftrightarrow E_j \subseteq E_{j_0},\ \text{for every } j \in J.$$

Proof. . □

Definition 191 *We call the* intersection *of* $\{E_j\}_{j\in J}$*, the event* $\bigcap_{j\in J} E_j$ *that occurs when all events of the family* $\{E_j\}_{j\in J}$ *occur. In symbols,*

$$\bigcap_{j\in J} E_j \stackrel{def}{=} \{\omega \in \Omega : \omega \in E_j\ \forall j \in J\}.$$

Proposition 192 *We have*

1. $\bigcap_{j\in J} E_j = E$ *for any family* $\{E_j\}_{j\in J}$ *of events of* Ω *such that* $E_j = E$*, for every* $j \in J$*;*
2. $\left(\bigcap_{j\in J_1} E_j\right) \cap \left(\bigcap_{j\in J_2} E_j\right) = \bigcap_{j\in J} E_j$*, for all* $J_1, J_2 \subseteq J$ *such that* $J_1 \cup J_2 = J$*;*
3. $\bigcap_{j\in J} E_j = \bigcap_{j\in J} E_{\pi(j)}$ *for any family* $\{E_j\}_{j\in J}$ *of events of* Ω *and any bijection* $\pi : J \to J$*.*

Proof. . □

Proposition 193 *We have*

$$\bigcap_{j\in J} E_j = E_{j_0},\ \text{for some } j_0 \in J,\ \Leftrightarrow E_{j_0} \subseteq E_j,\ \text{for every } j \in J.$$

Proof. . □

Definition 194 *We say that the family* $\{E_j\}_{j\in J}$ *is* jointly exhaustive *or* complete *if*

$$\bigcup_{j\in J} E_j = \Omega. \tag{4.2}$$

Definition 195 *We say that the events in* $\{E_j\}_{j\in J}$ *are* pairwise incompatible *if*

$$E_{j_1} \cap E_{j_2} = \varnothing,$$

for every $j_1, j_2 \in J$ *s.t.* $j_1 \neq j_2$

Definition 196 *We say that* $\{E_j\}_{j\in J}$ *is a* partition *of* Ω *if* $\{E_j\}_{j\in J}$ *is a jointly exhaustive family of pairwise incompatible events.*

Also the De Morgan and distributive laws can be extended to families of events of a sample space.

Proposition 197 (De Morgan laws) *We have*

1. $\left(\bigcup_{j\in J} E_j\right)^c = \bigcap_{j\in J} E_j^c$;
2. $\left(\bigcap_{j\in J} E_j\right)^c = \bigcup_{j\in J} E_j^c$.

Proof. . □

Proposition 198 (Distributive laws) *Let* $\{F_k\}_{k\in K}$ *be another family of events of* $\nleq$. *We have*

3. $\left(\bigcup_{j\in J} E_j\right) \cap \left(\bigcup_{k\in K} F_k\right) = \bigcup_{(j,k)\in J\times K}(E_j \cap F_k)$;
4. $\left(\bigcap_{j\in J} E_j\right) \cup \left(\bigcap_{k\in K} E_k\right) = \bigcap_{(j,k)\in J\times K}(E_j \cup F_k)$.

Proof. . □

In some circumstances it may be useful to consider union and intersection of families of events of a sample space. In these contexts, the rules to apply are rather natural.

Let Ω be a sample space and let $\{E_j\}_{j\in J}$ and $\{F_k\}_{k\in K}$ be families of events of Ω.

Proposition 199 *We have*

$$\left(\bigcup_{j\in J} E_j\right) \cup \left(\bigcup_{k\in K} F_k\right) = \bigcup_{(j,k)\in J\times K} (E_j \cup F_k)$$

and

$$\left(\bigcap_{j\in J} E_j\right) \cap \left(\bigcap_{k\in K} E_k\right) = \bigcap_{(j,k)\in J\times K} (E_j \cap F_k).$$

Proof. . □

4.1.2 Indicator Function of an Event

Let Ω be a sample space and let E be an event of Ω.

Definition 200 *We call the* indicator function *of* E *the function* $1_E : \Omega \to \mathbb{R}$, *more briefly* 1_E, *given by*

$$1_E(\omega) \stackrel{def}{=} \begin{cases} 1 & \text{if } \omega \in E \\ 0 & \text{if } \omega \notin E \end{cases}.$$

Let E and F events of Ω and let 1_E and 1_F their indicator functions.

Remark 201 *We have*

1. $1_E = 1_F \Leftrightarrow E = F$, *in particular* $1_E = 1 \Leftrightarrow E = \Omega$ *and* $1_E = 0 \Leftrightarrow E = \varnothing$;
2. $1_E \leq 1_F \Leftrightarrow E \subseteq F$;
3. $1_{E^c} = 1 - 1_E$;
4. $1_{E \cap F} = 1_E 1_F$,*in particular* $1_E 1_F = 0 \Leftrightarrow E \cap F = \varnothing$;
5. $1_{E \cup F} = 1_E + 1_F - 1_{E \cap F}$;
6. $1_{F-E} = 1_F - 1_F 1_E$;
7. $1_{E \Delta F} = 1_{F-E} + 1_{E-F}$.

Remark 202 *We have*

4.1.3 Cardinality of an Event

Let Ω be a sample space. For any $E \subseteq \Omega$ we write $\mathcal{P}(E)$ for the set of the parts (power set) of E.

Definition 203 *Following the standard terminology and notation of set theory, we refer to the number of the sample points in E as the* cardinality *of E, in symbols $|E|$.*

Some elementary properties of the cardinality of sets are presented in Appendix 19. Here, we just recall

Theorem 204 *Assume*

1. *E is finite and contains k sample point for some $k \in \mathbb{N}$, that is $|E| = k$. Then, $|\mathcal{P}(E)| = 2^k = 2^{|E|}$.*
2. *E is denumerable, that is $|E| = \aleph_0$, where $\aleph_0$, referred to as* aleph-zero *or* aleph-naught*, is the smallest infinite cardinal number. Then, $|\mathcal{P}(E)| = 2^{\aleph_0} = \aleph_1$, where $\aleph_1$, referred to as* aleph-one*, is the second smallest infinite cardinal number (see the Cantor continuum hypothesis).*
3. *E has the cardinality of the continuum, that is $|E| = \aleph_1$. Then, $|\mathcal{P}(E)| = 2^{\aleph_1} > \aleph_1$.*

Definition 205 *We say that E is* countable *if E is finite or denumerable.*

4.2 Families of Events

Given the sample space Ω of a random phenomenon, the power set $\mathcal{P}(\Omega)$ of Ω represents the family of all events whose occurrence might be in principle assessed by an observer of the phenomenon. In other words, $\mathcal{P}(\Omega)$ represents all information on the random phenomenon which might be observable. However, in many cases, such information might be too large to be mathematically handled in a non trivial way or the information which is actually observable is only a piece of all information: part of the information could be hidden or, as it happens in stochastic phenomena, it could be revealed only progressively in time. Therefore, it is necessary to consider suitable subfamilies of $\mathcal{P}(\Omega)$ which are closed under the logical manipulation of the contained events.

4.2.1 Algebras of Events

Let Ω be the sample space Ω of a random phenomenon, let $\mathcal{P}(\Omega)$ be the power set of Ω, and let $\mathcal{E}$ a nonempty subfamily of $\mathcal{P}(\Omega)$.

Definition 206 *We we say that $\mathcal{E}$ is an* algebra of events *of Ω if the following conditions are fulfilled:*

1. *$E^c \in \mathcal{E}$ for every $E \in \mathcal{E}$;*
2. *$E_1 \cup E_2 \in \mathcal{E}$ for all $E_1, E_2 \in \mathcal{E}$.*

Example 207 *The family $\mathcal{P}(\Omega)$ itself is an algebra of events of Ω.*

Definition 208 *We call $\mathcal{P}(\Omega)$ the* complete information *on Ω.*

The complete information represents the largest piece of information on a random phenomenon.

Example 209 *The family $\{\varnothing, \Omega\}$ is an algebra of events of Ω.*

Definition 210 *We call $\{\varnothing, \Omega\}$ the* trivial information *on Ω.*

The trivial information represents the smallest piece of information on a random phenomenon.

Example 211 *Assume the sample space Ω contains at least two distinct outcomes, say ω_0 and ω_1, and consider an event E_0 of Ω such that $\omega_0 \in E_0$ and $\omega_1 \notin E_0$, for instance, $E_0 \equiv \{\omega_0\}$. Then, the family*

$$\mathcal{E}_0 \equiv \{\varnothing, \Omega, E_0, E_0^c\}$$

is a non trivial algebra of events of Ω.

Definition 212 *We call the algebra $\mathcal{E}_0$ presented in Example 211 the* Bernoulli information *on Ω.*

Example 213 *The family*

$$\mathcal{E}_{el} \equiv \{\{\omega\} : \omega \in \Omega\}$$

of all elementary events of Ω is not an algebra of events of Ω.

Example 214 *The family*

$$\mathcal{E}_{fin} \equiv \{E \in \mathcal{P}(\Omega) : |E| < \aleph_0\}$$

of all finite events of Ω is an algebra of events of Ω if and only if Ω itself is finite. In this case, we have $\mathcal{E}_{fin} = \mathcal{P}(\Omega)$.

Example 215 *The family*

$$\mathcal{E}_{fin\text{-}cofin} \equiv \{E \in \mathcal{P}(\Omega) : |E| < \aleph_0 \ \vee \ |E^c| < \aleph_0\}$$

of all events of Ω which are finite or have finite complement is an algebra of events of Ω.

Discussion. Note first that the family $\mathcal{E}_{\text{fin-cofin}}$ is nonempty since $|\varnothing| = 0$, which implies $\varnothing \in \mathcal{E}_{\text{fin-cofin}}$.

Now, to show that 1 of Definition 206 holds true, consider a event $E \in \mathcal{E}_{\text{fin-cofin}}$. Two cases are possible:

1. $|E| < \aleph_0$;
2. $|E^c| < \aleph_0$.

In Case 1, since $(E^c)^c = E$, the complement of E^c has finite cardinality. This implies $E^c \in \mathcal{E}_{\text{fin-cofin}}$. In Case 2, the event E^c itself has finite cardinality, which again implies $E^c \in \mathcal{E}_{\text{fin-cofin}}$.

In the end, to show that 2 of Definition 206 holds true, consider two events $E_1, E_2 \in \mathcal{E}_{\text{fin-cofin}}$. Again two cases are possible:

1. $|E_j| < \aleph_0$ for $j = 1, 2$;
2. $|E_1^c| < \aleph_0$ or $|E_2^c| < \aleph_0$.

In Case 1, we clearly have

$$|E_1 \cup E_2| \leq |E_1| + |E_2| < \aleph_0,$$

which implies $E_1 \cup E_2 \in \mathcal{E}_{\text{fin-cofin}}$. In Case 2, we have

$$|(E_1 \cup E_2)^c| = |E_1^c \cap E_2^c| \leq \min\left\{|E_1^c|, |E_2^c|\right\} < \aleph_0.$$

Therefore, the complement $(E_1 \cup E_2)^c$ of $E_1 \cup E_2$ has finite cardinality. Hence, $E_1 \cup E_2 \in \mathcal{E}_{\text{fin-cofin}}$ and the argument is completed. □

Definition 216 *We call the algebra $\mathcal{E}_{\text{fin-cofin}}$ presented in Example 215 the* finite-cofinite information *on Ω.*

Remark 217 *We have*

$$\mathcal{E}_{\text{fin-cofin}} = \mathcal{P}(\Omega) \Leftrightarrow |\Omega| < \aleph_0.$$

Let $\mathcal{E}$ be an algebra of events of Ω.

Proposition 218 *From Definition 206 it follows*

1. *the events $\varnothing, \Omega \in \mathcal{E}$;*
2. *the event $\bigcup_{k=1}^n E_k \in \mathcal{E}$, for every finite sequence $(E_k)_{k=1}^n$ in $\mathcal{E}$;*
3. *the event $\bigcap_{k=1}^n E_k \in \mathcal{E}$, for every finite sequence $(E_k)_{k=1}^n$ in $\mathcal{E}$;*
4. *the event $F - E \in \mathcal{E}$, for all $E, F \in \mathcal{E}$.*

Proof. . □

It is useful to provide some basic rules for generating algebras of events from simple families of events of Ω.

Proposition 219 *Assume $\mathfrak{E}$ is any collection of algebras of events of Ω. Then, the family $\bigcap_{\mathcal{E}\in\mathfrak{E}}\mathcal{E}$ is an algebra of events of Ω.*

Proof. . □

Given a family $\mathcal{B}$ of events of Ω, write $\mathfrak{E}(\mathcal{B})$ for the collection of all the algebras of events of Ω containing $\mathcal{B}$.

Proposition 220 *The collection $\mathfrak{E}(\mathcal{B})$ is nonempty and the family $\bigcap_{\mathcal{E}\in\mathfrak{E}(\mathcal{B})}\mathcal{E}$ is an algebra of events of Ω containing $\mathcal{B}$. More specifically, $\bigcap_{\mathcal{E}\in\mathfrak{E}(\mathcal{B})}\mathcal{E}$ is the smallest algebra of events of Ω containing $\mathcal{B}$.*

Proof. . □

Definition 221 *We call the family $\bigcap_{\mathcal{E}\in\mathfrak{E}(\mathcal{B})}\mathcal{E}$ introduced in Proposition 220 the* algebra generated *by $\mathcal{B}$. In terms of notation, we write*

$$\bigcap_{\mathcal{E}\in\mathfrak{E}(\mathcal{B})}\mathcal{E}\equiv\alpha(\mathcal{B}).$$

Let $\mathcal{E}$ be an algebra of events of Ω.

Definition 222 *We say that a family $\mathcal{B}$ of events of Ω is a* basis *for $\mathcal{E}$ if $\alpha(\mathcal{B})=\mathcal{E}$.*

Remark 223 *Assume $\mathcal{B}$ itself is an algebra of events of Ω. Then, $\alpha(\mathcal{B})=\mathcal{B}$.*

Remark 224 *Assume $\mathcal{B}_1,\mathcal{B}_2$ are families of events of Ω such that $\mathcal{B}_1\subseteq\mathcal{B}_2$. Then, $\alpha(\mathcal{B}_1)\subseteq\alpha(\mathcal{B}_2)$.*

Example 225 *Assume the sample space Ω contains at least two distinct outcomes and consider an event E_0 of Ω such that $\varnothing\subset E_0\subset\Omega$, Then, we have*

$$\alpha(\{E_0\})=\mathcal{E}_0,$$

where $\mathcal{E}_0\equiv\{\varnothing,\Omega,E_0,E_0^c\}$ is the Bernoulli information (see Example 211).

Discussion. . □

Example 226 *The algebra of events of Ω generated by the family $\mathcal{E}_{el}$ of all elementary events of Ω is the finite-cofinite information (see Definition 216). In symbols,*

$$\alpha(\mathcal{E}_{el})=\mathcal{E}_{\textit{fin-cofin}}.$$

Hence, the family of all elementary events is a basis for the finite-cofinite information. In particular, if Ω is finite, Then, the algebra of events of Ω generated by $\mathcal{E}_{el}$ is the complete information.

Let $\mathcal{E}$ be an algebra of events of Ω and let $\mathcal{B}$ be a basis for $\mathcal{E}$.

Proposition 227 *Given any $\Omega_0 \subseteq \Omega$, the family of events*

$$\mathcal{E}_0 \equiv \{F \in \mathcal{P}(\Omega_0) : F = E \cap \Omega_0, \quad E \in \mathcal{E}\}$$

is an algebra of events of Ω_0. Furthermore, setting

$$\mathcal{B}_0 \equiv \{G \in \mathcal{P}(\Omega_0) : G = B \cap \Omega_0, \quad B \in \mathcal{B}\},$$

it turns out that $\mathcal{B}_0$ is a basis for $\mathcal{E}_0$.

Proof. To prove that $\mathcal{E}_0$ is an algebra of Ω_0, note first that $\mathcal{E}_0$ is non-empty since $\varnothing \in \mathcal{E}_0$. In fact, we can clearly write

$$\varnothing = \varnothing \cap \Omega_0, \quad \varnothing \in \mathcal{E}.$$

Alternatively, we can observe that $\Omega_0 \in \mathcal{E}$, since

$$\Omega_0 = \Omega \cap \Omega_0, \quad \Omega \in \mathcal{E}.$$

Second, assuming that $F \in \mathcal{E}_0$, that is $F = E \cap \Omega_0$ for some $E \in \mathcal{E}$, consider $F^c_{\Omega_0}$. We have

$$F^c_{\Omega_0} = (E \cap \Omega_0)^c_{\Omega_0} = E^c_{\Omega_0} \cup (\Omega_0)^c_{\Omega_0} = (E^c \cap \Omega_0) \cup \varnothing = E^c \cap \Omega_0, \quad E^c \in \mathcal{E}.$$

This implies $F^c_{\Omega_0} \in \mathcal{E}_0$. Third, assuming that $F_1, F_2 \in \mathcal{E}_0$, that is $F_k = E_k \cap \Omega_0$ for some $E_k \in \mathcal{E}$, on varying of $k = 1, 2$, consider $F_1 \cup F_2$. We have

$$F_1 \cup F_2 = (E_1 \cap \Omega_0) \cup (E_2 \cap \Omega_0) = (E_1 \cup E_2) \cap \Omega_0, \quad E_1 \cup E_2 \in \mathcal{E}.$$

It follows that $F_1 \cup F_2 \in \mathcal{E}_0$ and the proof is complete.

To prove that $\mathcal{B}_0$ is a basis for $\mathcal{E}_0$, we need to show that $\alpha(\mathcal{B}_0) = \mathcal{E}_0$, where $\alpha(\mathcal{B}_0)$ is the algebra generated by $\mathcal{B}_0$. The proof is achieved in two steps. First, we introduce the non-empty family of events of Ω.

$$\tilde{\mathcal{E}} \equiv \{E \in \mathcal{E} : E \cap \Omega_0 \in \alpha(\mathcal{B}_0)\},$$

and prove that

$$\tilde{\mathcal{E}} = \mathcal{E}. \tag{4.3}$$

Second, we use this result to show that

$$\alpha(\mathcal{B}_0) = \mathcal{E}_0. \tag{4.4}$$

To prove (4.3), we start to show that $\tilde{\mathcal{E}}$ is an algebra.

In fact, consider $E \in \tilde{\mathcal{E}}$, that is $E \cap \Omega_0 \in \alpha(\mathcal{B}_0)$, since we can write

$$E^c \cap \Omega_0 = (E^c \cap \Omega_0) \cup \varnothing = E^c_{\Omega_0} \cup (\Omega_0)^c_{\Omega_0} = (E \cap \Omega_0)^c_{\Omega_0},$$

and $\alpha(\mathcal{B}_0)$ is an algebra of events of Ω_0, it follows $E^c \cap \Omega_0 \in \alpha(\mathcal{B}_0)$, that is $E^c \in \tilde{\mathcal{E}}$. Moreover, if $(E_k)_{k=1}^n$ is a sequence of events in $\tilde{\mathcal{E}}$, that is $E_k \cap \Omega_0 \in \alpha(\mathcal{B}_0)$, for every $k = 1, \ldots, n$, since we have

$$(\textstyle\bigcup_{k=1}^n E_n) \cap \Omega_0 = \bigcup_{k=1}^n (E_n \cap \Omega_0),$$

and $\alpha(\mathcal{B}_0)$ is an algebra of events of Ω_0, we obtain $\bigcup_{k=1}^n E_n \in \tilde{\mathcal{E}}$, which completes the proof.

Now, we clearly have

$$\mathcal{B} \subseteq \tilde{\mathcal{E}} \subseteq \mathcal{E}. \tag{4.5}$$

In fact, by definition, $\tilde{\mathcal{E}} \subseteq \mathcal{E}$. Furthermore, if $B \in \mathcal{B}$, Then, $B \cap \Omega_0 \in \mathcal{B}_0 \subseteq \alpha(\mathcal{B}_0)$. We Then, have $B \in \tilde{\mathcal{E}}$, so that $\mathcal{B} \subseteq \tilde{\mathcal{E}}$.

As a consequence of (4.4),

$$\mathcal{E} = \alpha(\mathcal{B}) \subseteq \alpha\left(\tilde{\mathcal{E}}\right) = \tilde{\mathcal{E}} \subseteq \alpha(\mathcal{E}) = \mathcal{E}.$$

which yields the desired (4.3).

Finally, if $F \in \mathcal{E}_0$, that is $F = E \cap \Omega_0$ for some $E \in \mathcal{E}$, then, on account of (4.3), we have $E \in \tilde{\mathcal{E}}$, that is $E \cap \Omega_0 \in \alpha(\mathcal{B}_0)$. It follows

$$\mathcal{E}_0 \subseteq \alpha(\mathcal{B}_0). \tag{4.6}$$

On the other hand, by definition,

$$\mathcal{B}_0 \subseteq \mathcal{E}_0.$$

It follows,

$$\alpha(\mathcal{B}_0) \subseteq \alpha(\mathcal{E}_0) = \mathcal{E}_0. \tag{4.7}$$

Combining (4.6) and (4.7) we Then, obtain the desired (4.4). □

Definition 228 *We call the algebra $\mathcal{E}_0$ introduced in Proposition 227 the algebra of events of Ω_0* induced *by $\mathcal{E}$. We call $\mathcal{B}_0$ the basis of $\mathcal{E}_0$* induced *by $\mathcal{B}$.*

Proposition 229 *Let $n \in \mathbb{N}$ and let $\{E_k\}_{k=1}^n$ be a finite partition of Ω (see Definition 196). Set $\{1, \ldots, n\} \equiv N$ and consider the family $\tilde{\mathcal{E}}$ of events of Ω given by*

$$\mathcal{E} \stackrel{def}{=} \left\{ E \in \mathcal{P}(\Omega) : E = \bigcup_{k \in M} E_k, \quad M \in \mathcal{P}(N) \right\}.$$

We have

$$\mathcal{E} = \alpha(\{E_k\}_{k=1}^n). \tag{4.8}$$

Proof. Since $N \in \mathcal{P}(N)$ and $\bigcup_{k \in N} E_k = \Omega$, we have that $\Omega \in \tilde{\mathcal{E}} \neq \varnothing$. Alternatively, we can observe that since $\varnothing \in \mathcal{P}(N)$ and $\bigcup_{k \in \varnothing} E_k = \varnothing$ we have that $\varnothing \in \tilde{\mathcal{E}} \neq \varnothing$.

To prove that 1 of Definition 206 holds true, we consider an event $E \in \tilde{\mathcal{E}}$. We have $E = \bigcup_{k \in M_E} E_k$, for a suitable $M_E \in \mathcal{P}(N)$. Then,

$$E^c = \left(\bigcup_{k \in M_E} E_k \right)^c = \bigcap_{k \in M_E} E_k^c.$$

Now, since $\{E_k\}_{k=1}^n$ is a partition of Ω, we can write

$$\bigcap_{k \in M_E} E_k^c = \bigcup_{k \in N - M_E} E_k. \tag{4.9}$$

In fact, if a sample point $\omega \in \bigcap_{k \in M_E} E_k^c$, Then, $\omega \notin E_k$ for every $k \in M_E$. However, $\bigcup_{k \in N} E_n = \Omega$, Therefore, there exists $k_\omega \in N - M_E$ such that $\omega \in E_{k_\omega}$. This implies,

$$\bigcap_{k \in M_E} E_k^c \subseteq \bigcup_{k \in N - M_E} E_k.$$

Conversely, if $\omega \in \bigcup_{k\in N-M_E} E_k$, Then, there exists $k_\omega \in N - M_E$ such that $\omega \in E_{k_\omega}$, Since the events of $\{E_k\}_{k=1}^n$ are pairwise incompatible, this implies that $\omega \notin E_k$ for every $k \neq k_\omega$. A fortiori $\omega \notin E_k$ for every $k \in M_E$. Hence, $\omega \in E_k^c$ for every $k \in M_E$, which yields $\omega \in \bigcap_{k\in M_E} E_k^c$. We Then, have

$$\bigcup_{k\in N-M_E} E_n \subseteq \bigcap_{k\in M_E} E_k^c$$

and (4.9) follows.

As a consequence of (4.9), the family $\tilde{\mathcal{E}}$ is closed under complementation.

Similarly, to prove that 2 holds true, consider a couple of events $E, F \in \tilde{\mathcal{E}}$. We have $E = \bigcup_{k\in M_E} E_k$ and $F = \bigcup_{k\in M_F} E_k$, for suitable $M_E, M_F \in \mathcal{P}(N)$. It Then, follows

$$E \cup F = \bigcup_{k\in M_E\cup M_F} E_k. \tag{4.10}$$

In fact, since we clearly have

$$\bigcup_{k\in M_E} E_k \subseteq \bigcup_{k\in M_E\cup M_F} E_k \quad \text{and} \quad \bigcup_{k\in M_F} E_k \subseteq \bigcup_{k\in M_E\cup M_F} E_k$$

we obtain

$$E \cup F \subseteq \bigcup_{k\in M_E\cup M_F} E_k.$$

Conversely, given a sample point $\omega \in \bigcup_{k\in M_E\cup M_F} E_k$ there exists $k \in M_E$ or $k \in M_F$ such that $\omega \in E_k$. In the first case, $\omega \in E$; in the second case $\omega \in F$. Overall, $\omega \in E \cup F$, which shows that

$$\bigcup_{k\in M_E\cup M_F} E_k \subseteq E \cup F.$$

This completes the proof of (4.10).

As a consequence of (4.10), the family $\tilde{\mathcal{E}}$ is closed under finite union.

In light of what shown above, $\tilde{\mathcal{E}}$ turns out to be an algebra. In addition, we clearly have $\{E_k\}_{k=1}^n \subseteq \tilde{\mathcal{E}}$. Thanks to Remarks **??** and **??**, it Then, follows

$$\alpha(\{E_k\}_{k=1}^n) \subseteq \alpha\left(\tilde{\mathcal{E}}\right) = \tilde{\mathcal{E}}. \tag{4.11}$$

To prove the converse of (4.10), it is enough to observe that, according to the definition of $\tilde{\mathcal{E}}$ and (2) of Propositon 218, every $E \in \tilde{\mathcal{E}}$ necessarily belongs to $\alpha(\{E_k\}_{k=1}^n)$. This means that

$$\tilde{\mathcal{E}} \subseteq \alpha(\{E_k\}_{k=1}^n). \tag{4.12}$$

Combining (4.11) and (4.12), the desired (4.8) immediately follows. □

4.2.2 σ-Algebras of Events

Given a random phenomenon represented by a finite sample space $\Omega \equiv \{\omega_1, \ldots, \omega_n\}$, for some $n \in \mathbb{N}$, Then, every family of events of Ω is also finite. More precisely, the complete information $\mathcal{P}(\Omega)$ contains 2^n events and any other pieces of information, which is represented by an algebra of events of Ω, contains 2^k events for a suitable $k \in \{1, \ldots, n-1\}$. Therefore, we can manage

this case by elementary combinatorics. On the other hand, when the sample space Ω cannot be chosen finite, Then, any piece of information which aims to consider all elementary events has at least the power of the continuous. In symbols

$$|\Omega| \geq \aleph_0 \Rightarrow |\mathcal{P}(\Omega)| \geq 2^{\aleph_0} = \aleph_1.$$

In this case, combinatorics is no more applicable. Thus, it is necessary to extend the notion of algebra of events of Ω in a suitable way to use infinitesimal calculus.

Let Ω be a sample space and let $\mathcal{P}(\Omega)$ be the power set of Ω.

Definition 230 *We call a* σ-algebra *of events of* Ω *any nonempty subfamily* $\mathcal{E}$ *of* $\mathcal{P}(\Omega)$ *which has the following properties*

1. $E^c \in \mathcal{E}$ *for every* $E \in \mathcal{E}$;
2. $\bigcup_{n=1}^{\infty} E_n \in \mathcal{E}$ *for every sequence* $(E_n)_{n\geq 1}$ *in* $\mathcal{E}$.

Example 231 *The complete information* $\mathcal{P}(\Omega)$ *(see Example 207) is a* σ*-algebra of events of* Ω.

Remark 232 *Any finite algebra of events of* Ω *is a* σ*-algebra. In particular, the trivial information* $\{\varnothing, \Omega\}$ *is a* σ*-algebra.*

Remark 233 *If* Ω *is finite, Then, any algebra of events of* Ω *is a* σ*-algebra.*

Example 234 *If* Ω *is infinite, the finite-cofinite information* $\mathcal{E}_{fin\text{-}cofin}$ *(see Example 216) is not a* σ*-algebra.*

Example 235 *The family* $\mathcal{E}_{count} \equiv \{E \in \mathcal{P}(\Omega) : |E| \leq \aleph_0\}$ *of all countable events of* Ω *is a* σ*-algebra if and only if* Ω *itself is countable. In this case, we have* $\mathcal{E}_{count} = \mathcal{P}(\Omega)$.

Example 236 *The family* $\mathcal{E}_{count\text{-}cocount}$ *of all events of* Ω *which are countable or have countable complement, in symbols*

$$\mathcal{E}_{count\text{-}cocount} \equiv \{E \in \mathcal{P}(\Omega) : |E| \leq \aleph_0 \vee |E^c| \leq \aleph_0\},$$

is a σ*-algebra.*

Definition 237 *We call the algebra* $\mathcal{E}_{count\text{-}cocount}$ *introduced in Example 236 the* countable-cocountable information *on* Ω.

Remark 238 *We have*

$$\mathcal{E}_{count\text{-}cocount} = \mathcal{P}(\Omega) \Leftrightarrow |\Omega| = \aleph_0.$$

As a consequence of Definition 230 we have

Proposition 239 *Assume* $\mathcal{E}$ *is a* σ*-algebra of events of* Ω. *Then*

1. $\mathcal{E}$ *is a algebra of events of* Ω;
2. $\bigcap_{n=1}^{\infty} E_n \in \mathcal{E}$ *for every sequence* $(E_n)_{n\geq 1}$ *in* $\mathcal{E}$.

Proof. . □

Essentially the same basic rules provided for generating algebras from simple families of events of Ω apply in the case of σ-algebras.

Proposition 240 *Assume $\mathfrak{E}_\sigma$ is any collection of σ-algebras of events of Ω. Then, the family $\bigcap_{\mathcal{E}\in\mathfrak{E}_\sigma}\mathcal{E}$ is a σ-algebra.*

Proof. . □

Given a family $\mathcal{B}$ of events of Ω, write $\mathfrak{E}_\sigma(\mathcal{B})$ for the collection of all σ-algebras of events of Ω containing $\mathcal{B}$.

Proposition 241 *The collection $\mathfrak{E}_\sigma(\mathcal{B})$ is nonempty and the family $\bigcap_{\mathcal{E}\in\mathfrak{E}_\sigma(\mathcal{B})}\mathcal{E}$ is a σ-algebra of events of Ω containing $\mathcal{B}$. More specifically, $\bigcap_{\mathcal{E}\in\mathfrak{E}_\sigma(\mathcal{B})}\mathcal{E}$ is the smallest σ-algebra of events of Ω containing $\mathcal{B}$.*

Proof. . □

Definition 242 *We call the family $\bigcap_{\mathcal{E}\in\mathfrak{E}_\sigma(\mathcal{B})}\mathcal{E}$ introduced in Proposition 241 the σ-*algebra generated *by $\mathcal{B}$ and we use for it the simpler notation $\sigma(\mathcal{B})$.*

Let $\mathcal{E}$ be a σ- algebra of events of Ω

Definition 243 *We say that a family $\mathcal{B}$ of events of Ω is a* basis *for $\mathcal{E}$ if $\sigma(\mathcal{B})=\mathcal{E}$.*

Remark 244 *Assume a family $\mathcal{B}$ of events of Ω is itself a σ-algebra. Then, $\sigma(\mathcal{B})=\mathcal{B}$.*

Remark 245 *Assume $\mathcal{B}_1,\mathcal{B}_2$ are σ-algebras of events of Ω such that $\mathcal{B}_1\subseteq\mathcal{B}_2$. Then, $\sigma(\mathcal{B}_1)\subseteq\sigma(\mathcal{B}_2)$.*

Example 246 *The σ-algebra of events of Ω generated by the family $\mathcal{E}_{el}$ of all elementary events of Ω is the countable cocountable information (see Definition 237). In symbols,*

$$\sigma(\mathcal{E}_{el})=\mathcal{E}_{count\text{-}cocount}.$$

Hence, the family of all elementary events of Ω is a basis for the countable-cocountable information.

Remark 247 *We have*

$$\sigma(\mathcal{E}_{el})=\mathcal{P}(\Omega)\Leftrightarrow|\Omega|=\aleph_0.$$

Proposition 248 *Let $\mathcal{E}$ be a σ-algebra of events of Ω, let $\mathcal{B}$ be a basis for $\mathcal{E}$ and let $\Omega_0\subseteq\Omega$. Then, the family of events*

$$\mathcal{E}_0\equiv\{F\in\mathcal{P}(\Omega_0):F=E\cap\Omega_0,\quad E\in\mathcal{E}\}$$

is a σ-algebra of events of Ω_0. Furthermore, setting

$$\mathcal{B}_0\equiv\{C\in\mathcal{P}(\Omega_0):C=B\cap\Omega_0,\quad B\in\mathcal{B}\}$$

it turns out that $\mathcal{B}_0$ is a basis for $\mathcal{E}_0$.

Proof. Mutatis mutandis, the same proof of Proposition 227 applies. □

Definition 249 *We call the σ-algebra $\mathcal{E}_0$ introduced in Proposition 248 the σ-algebra of events of Ω_0* induced *by $\mathcal{E}$. We call the basis $\mathcal{B}_0$ the basis of $\mathcal{E}_0$* induced *by $\mathcal{B}$.*

Proposition 250 *Let $\mathcal{B} \equiv \{B_n\}_{n\geq 1}$ a countable partition of Ω (see Definition **??**). Consider the family $\mathcal{E}$ of events of Ω given by*

$$\mathcal{E} \stackrel{def}{=} \left\{ E \in \mathcal{P}(\Omega) : E = \bigcup_{n\in N} B_n, \quad N \in \mathcal{P}(\mathbb{N}) \right\}.$$

We have

$$\mathcal{E} = \sigma(\mathcal{B}).$$

Proof. Mutatis mutandis, the proof proceeds as that of Proposition 229. □

4.3 Probabilities

The next step towards a model for random phenomena is to introduce the notion of *probability* to rank the possibility of the occurrence of the events.

4.3.1 Empirical Probability

Assume to observe a random phenomenon which repeats in time under invariant conditions, for instance the flip of a fair coin, and assume that in n trials we record the occurrence of an event E, for instance the occurrence of heads, for k times.

Definition 251 *The number*

$$p_E \equiv \frac{k}{n}$$

is called the relative frequency of the event E.

Example 252 *The relative frequence of both outcomes* 1 *(representing heads) and* 0 *(representing tails) in a sequence of flips of a fair coin fluctuates around the value* $1/2$*, as the number n of the flips increases.*

Example 253 *The relative frequency of all outcomes* $1, \ldots, 6$ *in a sequence of rolls of a fair die fluctuates around the value* $1/6$*, as the number n of the rolls increases.*

Example 254 *The relative frequency of drawing j white balls in m draws with replacement from an urn containing N balls of which M are white and the remaining $N - M$ are black fluctuates around the value $\binom{m}{j} p^j q^{m-j}$, for $j = 0, 1, \ldots, m$, where $p \equiv \frac{M}{N}$ and $q \equiv \frac{N-M}{N} = 1 - p$, as the number n of the groups of m draws increases.*

Example 255 *The relative frequency of receiving k messages on a chat group of yours during a prescribed time interval Δt, say one day, fluctuates around the value $\frac{\lambda^k}{k!} e^{-\lambda}$, where λ stands for the average number of messages you have got in several past time intervals of the same length of Δt, as the number n of the prescribed time intervals increases.*

Remark 256 *We have*

1. $0 \leq p_E \leq 1$*, for any event E;*
2. $p_\Omega = 1$*;*
3. $p_{E \cup F} = p_E + p_F$*, for any pair E, F of incompatible events.*

Definition 257 *If the relative frequency of the occurrence of an event of a random phenomenon fluctuates around a well determined number, the latter is called the* empirical probability *of the event.*

4.3.2 Finitely Additive Probabilities

Let Ω be a sample space, and let $\mathcal{E}$ be an algebra of events of Ω.

Definition 258 *We say that a function $\mathbf{P} : \mathcal{E} \to \mathbb{R}_+$ is a* finitely additive probability *on Ω if it has the following properties*

1. $\mathbf{P}(\Omega) = 1$;
2. $\mathbf{P}(E \cup F) = \mathbf{P}(E) + \mathbf{P}(F)$ *for any pair E, F of incompatible events in $\mathcal{E}$.*

Definition 259 *If $\mathbf{P} : \mathcal{E} \to \mathbb{R}_+$ is a finitely additive probability on Ω, Then, for every $E \in \mathcal{E}$ the value $\mathbf{P}(E)$ is called the* probability *of the event E.*

Notation 260 *For any any outcome $\omega \in \Omega$ such that the elementary event $\{\omega\} \in \mathcal{E}$, we will follow the standard convention of writing $\mathbf{P}(\omega)$ rather than the cumbersome $\mathbf{P}(\{\omega\})$.*

Definition 261 *If $\mathbf{P} : \mathcal{E} \to \mathbb{R}_+$ is a finitely additive probability on Ω, we say that and event $E \in \mathcal{E}$ is* almost impossible *[resp.* almost sure*] if $E \neq \varnothing$ and $\mathbf{P}(E) = 0$ [resp. $E \neq \Omega$ and $\mathbf{P}(E) = 1$].*

Example 262 (Dirac probability) *For any fixed $\omega_0 \in \Omega$, the function $\mathbf{P}_0 : \mathcal{E} \to \mathbb{R}_+$ given by*

$$\mathbf{P}_0(E) \stackrel{def}{=} \begin{cases} 1 & \text{if } \omega_0 \in E \\ 0 & \text{if } \omega_0 \notin E \end{cases} \tag{4.13}$$

is a finitely additive probability on Ω such that

$$\mathbf{P}_0(\omega_0) = 1. \tag{4.14}$$

Discussion. The validity of 1 in Definition 258 and Equation (4.14) being evident, we just need to prove that 2 in Definition 258 holds true. To this, consider any pair $E, F \in \mathcal{E}$ such that $E \cap F = \varnothing$. Only two cases are possible

$$\omega_0 \in E \cup F \quad \text{or} \quad \omega_0 \notin E \cup F.$$

In the first case, we have

$$\mathbf{P}_0(E \cup F) = 1.$$

On the other hand, since E and F are incompatible, ω_0 has to be either in E or F. This implies

$$\mathbf{P}_0(E) + \mathbf{P}_0(F) = 1.$$

In the second case, we have

$$\mathbf{P}_0(E \cup F) = 0.$$

On the other hand, ω_0 can be neither in E nor F. Thus

$$\mathbf{P}_0(E) + \mathbf{P}_0(F) = 0.$$

Hence, in both the cases we have

$$\mathbf{P}_0(E \cup F) = \mathbf{P}_0(E) + \mathbf{P}_0(F),$$

which proves that also 2 in Definition 258 holds true. □

Definition 263 *We call the probability presented in Example 262 the* Dirac probability *on* Ω *concentrated at* ω_0.

Note that any event implied by the elementary event $\{\omega_0\}$ is almost sure and any event which is not implied by $\{\omega_0\}$ is almost impossible. Therefore, the Dirac probability is used to model any "random" experiment with a "non-random" outcome.

Example 264 (Dirac probability) *Consider an urn containing only a certain number of white balls. Draw a ball from the urn. The probability that the drawn ball is white [resp. black] ball is clearly* 1 *[resp.* 0*].*

Example 265 (Bernoulli probability) *Assume the sample space* Ω *contains at least two sample points, say* ω_0 *and* ω_1*. Chose any* $p \in (0,1)$ *and write* $q \equiv 1-p$*. Then, the function* $\mathbf{P} : \mathcal{P}(\Omega) \to \mathbb{R}_+$ *given by*

$$\mathbf{P}(E) \overset{def}{=} \begin{cases} 1 & \text{if } \omega_0 \in E \text{ and } \omega_1 \in E \\ p & \text{if } \omega_0 \notin E \text{ and } \omega_1 \in E \\ q & \text{if } \omega_0 \in E \text{ and } \omega_1 \notin E \\ 0 & \text{if } \omega_0 \notin E \text{ and } \omega_1 \notin E \end{cases} \tag{4.15}$$

is a finitely additive probability measure on Ω *such that*

$$\mathbf{P}(\omega_1) = p \qquad \text{and} \qquad \mathbf{P}(\omega_0) = q, \tag{4.16}$$

Discussion. Also in this case, the validity of 1. in Definition 258 and Equation (4.16) being evident, we need to prove that 2 in Definition 258 holds true. To this, consider any pair E, F of incompatible events in $\mathcal{P}(\Omega)$. Focusing on the event E, the following cases are possible:

1. $\omega_0 \in E$ and $\omega_1 \in E$;
2. $\omega_0 \in E$ and $\omega_1 \notin E$;
3. $\omega_0 \notin E$ and $\omega_1 \in E$;
4. $\omega_0 \notin E$ and $\omega_1 \notin E$;

In Case 1, we have $\omega_0, \omega_1 \in E \cup F$ and, since $E \cap F = \varnothing$, we have also $\omega_0 \notin F$ and $\omega_1 \notin E$. Applying Equation (4.15) to E, F, and $E \cup F$, we Then, have

$$\mathbf{P}(E) = 1, \quad \mathbf{P}(F) = 0, \quad \mathbf{P}(E \cup F) = 1.$$

Therefore 2 in Definition 258 is satisfied.

In Case 2, considering the condition $E \cap F = \varnothing$, we have only two possible subcases $\omega_1 \in F$ or $\omega_1 \notin F$. In the first subcase, we obtain $\omega_0, \omega_1 \in E \cup F$, and, applying (4.15) to E, F, and $E \cup F$, it follows

$$\mathbf{P}(E) = q, \quad \mathbf{P}(F) = p, \quad \mathbf{P}(E \cup F) = 1.$$

On the other hand, $p + q = 1$. Hence, 2 in Definition 258 is satisfied. In the second subcase, we obtain $\omega_0 \in E \cup F$ and $\omega_1 \notin E \cup F$. Thus,

$$\mathbf{P}(E) = q, \quad \mathbf{P}(F) = 0, \quad \mathbf{P}(E \cup F) = q.$$

Still, 2 in Definition 258 holds true.

Case 3 can be clearly dealt with an analogous argument of Case 2. Once more 2 in Definition 258 holds true.

In the end, in case 4, we have four possible subcases

4.1 $\omega_0, \omega_1 \in F$, which implies $\omega_0, \omega_1 \in E \cup F$, so that

$$\mathbf{P}(E) = 0, \quad \mathbf{P}(F) = 1, \quad \mathbf{P}(E \cup F) = 1;$$

4.2 $\omega_0 \in F$ and $\omega_1 \notin F$, which implies $\omega_0 \in E \cup F$ and $\omega_1 \notin E \cup F$, so that

$$\mathbf{P}(E) = 0, \quad \mathbf{P}(F) = q, \quad \mathbf{P}(E \cup F) = q;$$

4.3 $\omega_0 \notin F$ and $\omega_1 \in F$, which implies $\omega_0 \notin E \cup F$ and $\omega_1 \in E \cup F$, so that

$$\mathbf{P}(E) = 0, \quad \mathbf{P}(F) = p, \quad \mathbf{P}(E \cup F) = p;$$

4.4 $\omega_0, \omega_1 \notin F$, which implies $\omega_0, \omega_1 \notin E \cup F$, so that

$$\mathbf{P}(E) = 0, \quad \mathbf{P}(F) = 0, \quad \mathbf{P}(E \cup F) = 0.$$

Hence, in each subcase 2 in Definition 258 holds true. This shows that also in Case 4, the desired 2 in Definition 258 holds true and completes the proof. □

Definition 266 (Bernoulli probability) *We call the probability presented in Example 265 the* Bernoulli probability *with* success parameter p.

The Bernoulli probability is used to model any random phenomenon with only two distinguishable possible outcomes, of which one is considered a "success" and the other is considered a "failure".

Example 267 (naive probability) *Assume the sample space Ω contains a finite number of sample points, say $\Omega \equiv \{\omega_1, \ldots, \omega_n\}$ for some $n \in \mathbb{N}$, and let $\mathbf{P}_k : \mathcal{P}(\Omega) \to \mathbb{R}_+$ be the Dirac probability concentrated at ω_k, for $k = 1, \ldots, n$. Then, the function $\mathbf{P} : \mathcal{P}(\Omega) \to \mathbb{R}_+$ given by*

$$\mathbf{P}(E) \overset{def}{=} \frac{1}{n} \sum_{k=1}^{n} \mathbf{P}_k(E) \tag{4.17}$$

is a finitely additive probability on Ω such that

$$\mathbf{P}(\omega_k) = \frac{1}{n}, \tag{4.18}$$

for every $k = 1, \ldots, n$.

Discussion. Equation (4.21) being evident, we only need to prove that Properties 1 and 2 hold true. Since $\mathbf{P}_k : \mathcal{P}(\Omega) \to \mathbb{R}_+$ is a probability, for every $k = 1, \ldots, n$, we have

$$\mathbf{P}(\Omega) = \frac{1}{n} \sum_{k=1}^{n} \mathbf{P}_k(\Omega) = \frac{1}{n} \sum_{k=1}^{n} 1 = \frac{1}{n} n = 1.$$

In addition, thanks to the properties of finite sums,

$$\begin{aligned} \mathbf{P}(E \cup F) &= \frac{1}{n} \sum_{k=1}^{n} \mathbf{P}_k(E \cup F) = \frac{1}{n} \sum_{k=1}^{n} (\mathbf{P}_k(E) + \mathbf{P}_k(F)) \\ &= \frac{1}{n} \sum_{k=1}^{n} \mathbf{P}_k(E) + \frac{1}{n} \sum_{k=1}^{n} \mathbf{P}_k(F) \\ &= \mathbf{P}(E) + \mathbf{P}(F), \end{aligned}$$

for any pair $E, F \in \mathcal{P}(\Omega)$ such that $E \cap F = \varnothing$. This completes the argument. □

Definition 268 (naive probability) *We call the probability presented in Example 267 the* naive *or* discrete uniform probability*.*

The naive probability is used to model any random phenomenon with a finite number of possible outcomes, all of which have the same possibility of realization.

Proposition 269 (naive probability) *We have*

$$\mathbf{P}(E) = \frac{|E|}{|\Omega|}, \tag{4.19}$$

for every event $E \in \mathcal{P}(\Omega)$, where $|E|$ [resp. $|\Omega|$] is the cardinality of E [resp. Ω].

Proof. We can write

$$E = \bigcup_{j \in \{1,\ldots,n \,:\, \omega_j \in E\}} \{\omega_j\},$$

where $\{\omega_{k_1}\}$ and $\{\omega_{k_2}\}$ are incompatible, for all $k_1, k_2 \in \{1, \ldots, n\}$ such that $k_1 \neq k_2$. As a consequence

$$\mathbf{P}(E) = \mathbf{P}\left(\bigcup_{j \in \{1,\ldots,n \,:\, \omega_j \in E\}} \{\omega_j\} \right) = \sum_{j \in \{1,\ldots,n \,:\, \omega_j \in E\}} \mathbf{P}(\omega_j).$$

Now, since $\mathbf{P}_k$ is the Dirac probability concentrated at ω_k, for $k = 1, \ldots, n$, we have

$$\mathbf{P}_k(\omega_j) = \begin{cases} 1, & \text{if } j = k, \\ 0, & \text{if } j \neq k, \end{cases}$$

for every $j = 1, \ldots, n$. It Then, follows,

$$\mathbf{P}(\omega_j) = \frac{1}{n} \sum_{k=1}^{n} \mathbf{P}_k(\omega_j) = \frac{1}{n} \left(\sum_{\substack{k=1 \\ k \neq j}}^{n} \mathbf{P}_k(\omega_j) + \mathbf{P}_j(\omega_j) \right) = \frac{1}{n}.$$

Therefore,

$$\mathbf{P}(E) = \sum_{j \in \{1,\ldots,n \,:\, \omega_j \in E\}} \frac{1}{n} = \frac{\sum_{j \in \{1,\ldots,n \,:\, \omega_j \in E\}} 1}{n} = \frac{|E|}{|\Omega|},$$

as claimed. □

Since $|E|$ [resp. $|\Omega|$] is the number of the sample points contained in E [resp. Ω], the naive probability of any event E of a random phenomenon expresses the ratio between the number of outcomes whose occurrence implies the occurrence of E, so called *favorable outcomes*, and the number of all *possible outcomes*. It is worth of mention that at the early stages of the probability theory, scholars wished to handle any kind of problem by means of a naive idea of probability very close to the one introduced in Definition 268. For instance, the following definition was widely used.

Definition 270 (early definition) *The probability of an event E of a random phenomenon is the ratio between the number of all favorable outcomes and the number of all possible outcomes, provided the latter are equally probable.*

However, this approach soon revealed to be unnecessarily involved and artificial. Nowdays, the circular[1] Definition 270 is no longer considered as a true definition, but only a way to compute the probabilities in case of finite sample spaces whose outcomes may be thought of having the same probability of occurrence by virtue of some symmetry reason.

Example 271 (binomial probability) *Assume the sample space Ω contains a finite number of sample points, say $\Omega \equiv \{\omega_0, \omega_1, \ldots, \omega_n\}$ for some $n \in \mathbb{N}$. Choose any $p \in (0,1)$ and set $q \equiv 1-p$. Then, the function $\mathbf{P} : \mathcal{P}(\Omega) \to \mathbb{R}_+$ given by*

$$\mathbf{P}(E) \overset{def}{=} \sum_{\{k \in \{0,1,\ldots,n\} \,:\, \omega_k \in E\}} \binom{n}{k} p^k q^{n-k}. \tag{4.20}$$

is a finitely additive probability measure on Ω, such that

$$\mathbf{P}(\omega_k) = \binom{n}{k} p^k q^{n-k}, \tag{4.21}$$

for every $k = 0, 1, \ldots, n$.

Discussion. Equation (4.21) being evident, we only need to prove that Properties 1 and 2 hold true. Applying Equation (4.20) to Ω and considering Newton's binomial formula, we obtain

$$\mathbf{P}(\Omega) = \sum_{\{k \in \{0,1,\ldots,n\} \,:\, \omega_k \in \Omega\}} \binom{n}{k} p^k q^{n-k} = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} = (p+q)^n = 1. \tag{4.22}$$

Now, given any $E, F \in \mathcal{P}(\Omega)$ such that $E \cap F = \varnothing$, we clearly have

$$\{k \in \{0,1,\ldots,n\} \,:\, \omega_k \in E\} \cap \{k \in \{0,1,\ldots,n\} \,:\, \omega_k \in F\} = \varnothing$$

and

$$\{k \in \{0,1,\ldots,n\} \,:\, \omega_k \in E \cup F\} = \{k \in \{0,1,\ldots,n\} \,:\, \omega_k \in E\} \cup \{k \in \{0,1,\ldots,n\} \,:\, \omega_k \in F\}.$$

Therefore, Applying Equation (4.20) to $E \cup F$, we obtain

$$\begin{aligned}
\mathbf{P}(E \cup F) &= \sum_{\{k \in \{0,1,\ldots,n\} \,:\, \omega_k \in E \cup F\}} \binom{n}{k} p^k q^{n-k} \\
&= \sum_{\{k \in \{0,1,\ldots,n\} \,:\, \omega_k \in E\} \cup \{k \in \{0,1,\ldots,n\} \,:\, \omega_k \in F\}} \binom{n}{k} p^k q^{n-k} \\
&= \sum_{\{k \in \{0,1,\ldots,n\} \,:\, \omega_k \in E\}} \binom{n}{k} p^k q^{n-k} + \sum_{\{k \in \{0,1,\ldots,n\} \,:\, \omega_k \in F\}} \binom{n}{k} p^k q^{n-k} \\
&= \mathbf{P}(E) + \mathbf{P}(F)
\end{aligned}$$

This shows that also Property 2 is satisfied. □

[1] How may we check that all possible outcomes are equally probable if we have not definition of probability yet?

Definition 272 (binomial probability) *We call the probability introduced in Example 271 the* binomial probability *on Ω with* number of trials *[resp.* success*]* parameter *n [resp. p].*

Example 273 (bivariate hypergeometric probability) *Assume the sample space Ω contains a finite number of sample points, say $\Omega \equiv \{\omega_0, \omega_1, \ldots, \omega_n\}$ for some $n \in \mathbb{N}$. Let $N, M \in \mathbb{N}$ such that $n \leq \min\{M, N-M\}$. Then, the function $\mathbf{P} : \mathcal{P}(\Omega) \to \mathbb{R}_+$ given by*

$$\mathbf{P}(E) \stackrel{def}{=} \sum_{\{k \in \{0,1,\ldots,n\} \,:\, \omega_k \in E\}} \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}, \quad \forall k = 0, 1, \ldots, n, \tag{4.23}$$

is a finitely additive probability measure on Ω, such that

$$\mathbf{P}(\omega_k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}, \tag{4.24}$$

for every $k = 0, 1, \ldots, n$.

Discussion. Equation (4.24) being evident, we only need to prove that Properties 1 and 2 hold true. We have

$$\mathbf{P}(\Omega) = \sum_{\{k \in \{0,1,\ldots,n\} \,:\, \omega_k \in \Omega\}} \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} = \sum_{k=0}^{n} \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}. \tag{4.25}$$

Therefore, to show that $\mathbf{P}(\Omega) = 1$ we can equivalently show that

$$\sum_{k=0}^{n} \binom{M}{k}\binom{N-M}{n-k} = \binom{N}{n}. \tag{4.26}$$

The latter is referred to as *Vandermonde's identity*. To show Equation (4.26), consider the partition $\{1, \ldots, M\}$ and $\{M+1, \ldots, N\}$ of the set $\{1, \ldots, N\}$ it is rather evident that all and only the subsets of $\{1, \ldots, N\}$ containing n elements can be composed by a subset of $\{1, \ldots, M\}$ containing k elements and a subset of $\{M+1, \ldots, N\}$ containing $n-k$ elements, on varying of $k = 0, 1, \ldots, n$. Therefore,

$$\begin{aligned} &|\{A \subseteq \{1, \ldots, N\} : |A| = n\}| \\ &= \sum_{k=0}^{n} |\{B \subseteq \{1, \ldots, M\} : |B| = k\}| \, |\{C \subseteq \{M+1, \ldots, N\} : |C| = n-k\}|. \end{aligned} \tag{4.27}$$

On the other hand,

$$|\{A \subseteq \{1, \ldots, N\} : |A| = n\}| = \binom{N}{n}, \tag{4.28}$$

$$|\{B \subseteq \{1, \ldots, M\} : |B| = k\}| = \binom{M}{k}, \tag{4.29}$$

and

$$|\{C \subseteq \{M+1, \ldots, N\} : |C| = n-k\}| = |\{C \subseteq \{1, \ldots, N-M\} : |C| = n-k\}| = \binom{N-M}{n-k}. \tag{4.30}$$

Combining (4.28)-(4.30) with (4.27), we obtain the desired (4.26). In the end, to show also Property 2 is satisfied, we can apply the same argument presented in the proof of Proposition 271. □

Definition 274 (bivariate hypergeometric probability) *We call the probability introduced in Example 273 the* bivariate hypergeometric probability *on* Ω *with* population size *[resp.* population success] parameter N *[resp.* M*] and* number of trials parameter n*.*

Let $\mathbf{P} : \mathcal{E} \to \mathbb{R}_+$ be an additive probability on Ω.

Proposition 275 *We have*

1. $\mathbf{P}(\varnothing) = 0$*;*
2. $\mathbf{P}(E^c) = 1 - \mathbf{P}(E)$ *for every* $E \in \mathcal{E}$*;*
3. $\mathbf{P}(F - E) = \mathbf{P}(F) - \mathbf{P}(E \cap F)$ *for all* $E, F \in \mathcal{E}$*, in particular* $\mathbf{P}(F - E) = \mathbf{P}(F) - \mathbf{P}(E)$ *when* $E \subseteq F$*;*
4. $\mathbf{P}(E) \leq \mathbf{P}(F)$ *for all* $E, F \in \mathcal{E}$ *such that* $E \subseteq F$*;*
5. $\mathbf{P}(E) \leq 1$ *for every* $E \in \mathcal{E}$*;*
6. $\mathbf{P}(E \cup F) = \mathbf{P}(E) + \mathbf{P}(F) - \mathbf{P}(E \cap F)$ *for all* $E, F \in \mathcal{E}$*, in particular* $\mathbf{P}(E \cup F) \leq \mathbf{P}(E) + \mathbf{P}(F)$*;*
7. $\mathbf{P}\left(\bigcup_{k=1}^{n} E_k\right) = \sum_{k=1}^{n} \mathbf{P}(E_k)$ *for every finite sequence* $(E_k)_{k=1}^{n}$ *of pairwise incompatible events in* $\mathcal{E}$*;*
8. $\mathbf{P}\left(\bigcup_{k=1}^{n} E_n\right) \leq \sum_{k=1}^{n} \mathbf{P}(E_n)$ *for every finite sequence* $(E_k)_{k=1}^{n}$ *in* $\mathcal{E}$*.*

Proof.

1. We have
$$\Omega \cup \varnothing = \Omega \quad \text{and} \quad \Omega \cap \varnothing = \varnothing.$$
Hence, we are in a position to apply the additive property of $\mathbf{P}$. We Then, obtain
$$\mathbf{P}(\Omega) + \mathbf{P}(\varnothing) = \mathbf{P}(\Omega).$$
Property 1 clearly follows.

2. For any $E \in \mathcal{E}$, we can write
$$E \cup E^c = \Omega \quad \text{and} \quad E \cap E^c = \varnothing.$$
Again, we are in a position to apply the additive property of $\mathbf{P}$. Considering 1, we obtain
$$\mathbf{P}(E) + \mathbf{P}(E^c) = \mathbf{P}(\Omega) = 1.$$
This immediately impliese the desired Property 2.

3. For all $E, F \in \mathcal{E}$, we can write
$$F = (F - E) \cup (F \cap E) \quad \text{and} \quad (F - E) \cap (F \cap E) = \varnothing.$$
Still from the additive property of $\mathbf{P}$, we obtain
$$\mathbf{P}(F) = \mathbf{P}(F - E) + \mathbf{P}(F \cap E).$$
Therefore, Property 3 holds true.

4. For all $E, F \in \mathcal{E}$ such that $E \subseteq F$, we have

$$\mathbf{P}(F) = \mathbf{P}(F - E) + \mathbf{P}(E). \tag{4.31}$$

On the other hand,

$$\mathbf{P}(F - E) \geq 0. \tag{4.32}$$

Equations (4.31) and (4.32) clearly imply Property 4.

5. For any $E \in \mathcal{E}$ we have

$$E \subseteq \Omega.$$

By virtue of Property 4 it Then, follows

$$\mathbf{P}(E) \leq \mathbf{P}(\Omega) = 1,$$

as desired.

6. We know that for all $E_1, E_2 \in \mathcal{E}$ such that $E_1 \cap E_2 = \varnothing$ we have

$$\mathbf{P}(E_1 \cup E_2) = \mathbf{P}(E_1) + \mathbf{P}(E_2).$$

Assume given some $n > 2$, we have

$$\mathbf{P}\left(\bigcup_{k=1}^{n} E_k\right) = \sum_{k=1}^{n} \mathbf{P}(E_k),$$

for any sequence $(E_k)_{k=1}^{n}$ in $\mathcal{E}$ of n pairwise incompatible events, and consider a sequence $(E_k)_{k=1}^{n+1}$ in $\mathcal{E}$ of $n+1$ pairwise incompatible events. For the associative property of the union of events, we can write

$$\bigcup_{k=1}^{n+1} E_k = \left(\bigcup_{k=1}^{n} E_k\right) \cup E_{n+1}.$$

On the other hand, since the events of the sequence $(E_k)_{k=1}^{n+1}$ are pairwise incompatible, also the events $\bigcup_{k=1}^{n} E_k$ and E_{n+1} are pairwise incompatible. In fact, on account of the distributive property of the intersection with respect the union, we have

$$\left(\bigcup_{k=1}^{n} E_k\right) \cap E_{n+1} = \bigcup_{k=1}^{n} (E_k \cap E_{n+1}) = \bigcup_{k=1}^{n} \varnothing = \varnothing.$$

Therefore, we can write

$$\mathbf{P}\left(\bigcup_{k=1}^{n+1} E_k\right) = \mathbf{P}\left(\left(\bigcup_{k=1}^{n} E_k\right) \cup E_{n+1}\right) = \mathbf{P}\left(\bigcup_{k=1}^{n} E_k\right) + \mathbf{P}(E_{n+1}).$$

By virtue of the inductive hypothesis and the associativity of finite sums it follows

$$\mathbf{P}\left(\bigcup_{k=1}^{n+1} E_k\right) = \sum_{k=1}^{n} \mathbf{P}(E_k) + \mathbf{P}(E_{n+1}) = \sum_{k=1}^{n+1} \mathbf{P}(E_k).$$

This completes the proof.

7. For all $E, F \in \mathcal{E}$ we can write

$$E \cup F = (E - F) \cup (F - E) \cup (E \cap F),$$

where the events $E - F$, $F - E$, and $E \cap F$ are pairwise incompatible. Therefore,

$$\mathbf{P}(E \cup F) = \mathbf{P}(E - F) + \mathbf{P}(F - E) + \mathbf{P}(E \cap F). \tag{4.33}$$

On the other hand,

$$\mathbf{P}(E - F) = \mathbf{P}(E) - \mathbf{P}(E \cap F) \quad \text{and} \quad \mathbf{P}(F - E) = \mathbf{P}(F) - \mathbf{P}(E \cap F). \tag{4.34}$$

Hence, replacing (4.33) in Equation (4.34), we obtain

$$\begin{aligned}\mathbf{P}(E \cup F) &= \mathbf{P}(E) - \mathbf{P}(E \cap F) + \mathbf{P}(F) - \mathbf{P}(E \cap F) + \mathbf{P}(E \cap F) \\ &= \mathbf{P}(E) + \mathbf{P}(F) - \mathbf{P}(E \cap F),\end{aligned}$$

as desired. In particular, the positivity of $\mathbf{P}$ implies

$$\mathbf{P}(E \cup F) \leq \mathbf{P}(E) + \mathbf{P}(F).$$

8. As a consequence of Property 6, we know that for all $E_1, E_2 \in \mathcal{E}$ we have

$$\mathbf{P}(E_1 \cup E_2) \leq \mathbf{P}(E_1) + \mathbf{P}(E_2).$$

Assume given some $n > 2$, we have

$$\mathbf{P}\left(\bigcup_{k=1}^{n} E_k\right) \leq \sum_{k=1}^{n} \mathbf{P}(E_k),$$

for any sequence $(E_k)_{k=1}^{n}$ in $\mathcal{E}$ of n events and consider a sequence $(E_k)_{k=1}^{n+1}$ in $\mathcal{E}$ of $n+1$ events. An analogous argument to that used in the proof of Property 7 yields

$$\begin{aligned}\mathbf{P}\left(\bigcup_{k=1}^{n+1} E_k\right) = \mathbf{P}\left(\bigcup_{k=1}^{n} E_k \cup E_{n+1}\right) &\leq \mathbf{P}\left(\bigcup_{k=1}^{n} E_k\right) + \mathbf{P}(E_{n+1}) \\ &= \sum_{k=1}^{n} \mathbf{P}(E_k) + \mathbf{P}(E_{n+1}) = \sum_{k=1}^{n+1} \mathbf{P}(E_k).\end{aligned}$$

This completes the proof. □

Proposition 276 (Bonferroni Inequalities) *We have*

$$\mathbf{P}\left(\textstyle\bigcup_{k=1}^{n} E_k\right) \geq \displaystyle\sum_{k=1}^{n} \mathbf{P}(E_k) - \sum_{\substack{k,\ell=1 \\ \ell>k}}^{n} \mathbf{P}(E_k \cap E_\ell) \tag{4.35}$$

and

$$\mathbf{P}\left(\textstyle\bigcup_{k=1}^{n} E_k\right) \leq \displaystyle\sum_{k=1}^{n} \mathbf{P}(E_k) - \sum_{\substack{k,\ell=1 \\ \ell>k}}^{n} \mathbf{P}(E_k \cap E_\ell) + \sum_{\substack{k,\ell,m=1 \\ m>\ell>k}}^{n} \mathbf{P}(E_k \cap E_\ell \cap E_m), \tag{4.36}$$

for any finite sequence $(E_k)_{k=1}^{n}$ *in* $\mathcal{E}$.

Proof. With regard to Equation (4.35), observe that we clearly have

$$E_k - \bigcup_{\substack{\ell=1 \\ \ell>k}}^{n} (E_k \cap E_\ell) \subseteq E_k$$

for every $k = 1, \ldots, n$. Hence,

$$\bigcup_{k=1}^{n} \left(E_k - \bigcup_{\substack{\ell=1 \\ \ell>k}}^{n} (E_k \cap E_\ell) \right) \subseteq \bigcup_{k=1}^{n} E_k$$

As a consequence, on account of 4 in Proposition 275, we have

$$\mathbf{P}\left(\bigcup_{k=1}^{n} E_k\right) \geq \mathbf{P}\left(\bigcup_{k=1}^{n} \left(E_k - \bigcup_{\substack{\ell=1 \\ \ell>k}}^{n} (E_k \cap E_\ell) \right) \right) \tag{4.37}$$

Now, we have

$$\left(E_{k_1} - \bigcup_{\substack{\ell=1 \\ \ell>k_1}}^{n} (E_k \cap E_\ell) \right) \cap \left(E_{k_2} - \bigcup_{\substack{\ell=1 \\ \ell>k_2}}^{n} (E_k \cap E_\ell) \right) = \varnothing \tag{4.38}$$

for all $k_1, k_2 \in \{1, \ldots, n\}$ such that $k_1 \neq k_2$. In fact, assume $k_1 < k_2$ and consider

$$\omega \in E_{k_1} - \bigcup_{\substack{\ell=1 \\ \ell>k_1}}^{n} (E_{k_1} \cap E_\ell)\,.$$

Then, $\omega \in E_{k_1}$ and $\omega \notin E_\ell$ for every $\ell \in \{1, \ldots, n\}$ such that $\ell > k_1$. If it were not so, there would be $\ell_\omega > k_1$ such that $\omega \in E_{\ell_\omega}$. Hence, $\omega \in E_{k_1} \cap E_{\ell_\omega}$. Thus, $\omega \in \bigcup_{\substack{\ell=1 \\ \ell>k_1}}^{n} (E_{k_1} \cap E_\ell)$ and $\omega \notin E_{k_1} - \bigcup_{\substack{\ell=1 \\ \ell>k_1}}^{n} (E_{k_1} \cap E_\ell)$. In particular, $\omega \notin E_{k_2}$ which implies $\omega \notin E_{k_2} - \bigcup_{\substack{\ell=1 \\ \ell>k_2}}^{n} (E_{k_2} \cap E_\ell)$. This proves (4.38). As a consequence, considering 7 in Proposition 275, we have

$$\mathbf{P}\left(\bigcup_{k=1}^{n} \left(E_k - \bigcup_{\substack{\ell=1 \\ \ell>k}}^{n} (E_k \cap E_\ell) \right) \right) = \sum_{k=1}^{n} \mathbf{P}\left(E_k - \bigcup_{\substack{\ell=1 \\ \ell>k}}^{n} (E_k \cap E_\ell) \right). \tag{4.39}$$

On the other hand, we also clearly have

$$\bigcup_{\substack{\ell=1 \\ \ell>k}}^{n} (E_k \cap E_\ell) \subseteq E_k,$$

for every $k = 1, \ldots, n$. Hence, thanks to 3 and 8 in Proposition 275, we obtain

$$\mathbf{P}\left(E_k - \bigcup_{\substack{\ell=1 \\ \ell>k}}^{n} (E_k \cap E_\ell) \right) = \mathbf{P}(E_k) - \mathbf{P}\left(\bigcup_{\substack{\ell=1 \\ \ell>k}}^{n} (E_k \cap E_\ell) \right) \geq \mathbf{P}(E_k) - \sum_{\substack{\ell=1 \\ \ell>k}}^{n} \mathbf{P}(E_k \cap E_\ell)\,. \tag{4.40}$$

In the end, combining (4.37), (4.39) and (4.40), it follows

$$\mathbf{P}\left(\bigcup_{k=1}^{n} E_k\right) \geq \sum_{k=1}^{n} \left(\mathbf{P}\left(E_k\right) - \sum_{\substack{\ell=1 \\ \ell>k}}^{n} \mathbf{P}\left(E_k \cap E_\ell\right) \right) = \sum_{k=1}^{n} \mathbf{P}\left(E_k\right) - \sum_{k=1}^{n} \sum_{\substack{\ell=1 \\ \ell>k}}^{n} \mathbf{P}\left(E_k \cap E_\ell\right),$$

which is the desired (4.35). With regard to Equation (4.36), we observe that, applying Equation (4.35) to $\bigcup_{\substack{\ell=1 \\ \ell>k}}^{n} \left(E_k \cap E_\ell\right)$, we obtain

$$\mathbf{P}\left(\bigcup_{\substack{\ell=1 \\ \ell>k}}^{n} \left(E_k \cap E_\ell\right)\right) \geq \sum_{\substack{\ell=1 \\ \ell>k}}^{n} \mathbf{P}\left(E_k \cap E_\ell\right) - \sum_{\substack{\ell,m=1 \\ m>\ell>k}}^{n} \mathbf{P}\left(E_k \cap E_\ell \cap E_m\right) \tag{4.41}$$

for every $k = 1, \ldots, n$. Now, since

$$E_k = \left(E_k - \bigcup_{\substack{\ell=1 \\ \ell>k}}^{n} \left(E_k \cap E_\ell\right) \right) \cup \bigcup_{\substack{\ell=1 \\ \ell>k}}^{n} \left(E_k \cap E_\ell\right),$$

we have

$$\bigcup_{k=1}^{n} E_k = \bigcup_{k=1}^{n} \left(\left(E_k - \bigcup_{\substack{\ell=1 \\ \ell>k}}^{n} \left(E_k \cap E_\ell\right) \right) \cup \bigcup_{\substack{\ell=1 \\ \ell>k}}^{n} \left(E_k \cap E_\ell\right) \right),$$

where

$$\left(E_k - \bigcup_{\substack{\ell=1 \\ \ell>k}}^{n} \left(E_k \cap E_\ell\right) \right) \cap \bigcup_{\substack{\ell=1 \\ \ell>k}}^{n} \left(E_k \cap E_\ell\right) = \varnothing.$$

Therefore, thanks to 8 of Proposition 275 and considering 4.41, it follows

$$\begin{aligned}
\mathbf{P}\left(\bigcup_{k=1}^{n} E_k\right) &= \mathbf{P}\left(\bigcup_{k=1}^{n} \left(\left(E_k - \bigcup_{\substack{\ell=1 \\ \ell>k}}^{n} \left(E_k \cap E_\ell\right) \right) \cup \bigcup_{\substack{\ell=1 \\ \ell>k}}^{n} \left(E_k \cap E_\ell\right) \right) \right) \\
&\leq \sum_{k=1}^{n} \mathbf{P}\left(\left(E_k - \bigcup_{\substack{\ell=1 \\ \ell>k}}^{n} \left(E_k \cap E_\ell\right) \right) \cup \bigcup_{\substack{\ell=1 \\ \ell>k}}^{n} \left(E_k \cap E_\ell\right) \right) \\
&= \sum_{k=1}^{n} \mathbf{P}\left(E_k - \bigcup_{\substack{\ell=1 \\ \ell>k}}^{n} \left(E_k \cap E_\ell\right) \right) + \mathbf{P}\left(\bigcup_{\substack{\ell=1 \\ \ell>k}}^{n} \left(E_k \cap E_\ell\right) \right) \\
&\leq \sum_{k=1}^{n} \mathbf{P}\left(E_k\right) + \mathbf{P}\left(\bigcup_{\substack{\ell=1 \\ \ell>k}}^{n} \left(E_k \cap E_\ell\right) \right) \\
&\leq \sum_{k=1}^{n} \mathbf{P}\left(E_k\right) - \sum_{\substack{\ell=1 \\ \ell>k}}^{n} \mathbf{P}\left(E_k \cap E_\ell\right) + \sum_{\substack{\ell,m=1 \\ m>\ell>k}}^{n} \mathbf{P}\left(E_k \cap E_\ell \cap E_m\right),
\end{aligned}$$

as desired. □

Proposition 277 *We have*

$$\mathbf{P}\left(\bigcup_{k=1}^{n} E_k\right) = \sum_{k=1}^{n} \mathbf{P}\left(E_k\right) - \sum_{\substack{k,\ell=1 \\ \ell>k}}^{n} \mathbf{P}\left(E_k \cap E_\ell\right) + \sum_{\substack{k,\ell,m=1 \\ m>\ell>k}}^{n} \mathbf{P}\left(E_k \cap E_\ell \cap E_m\right) - \cdots + (-1)^{n+1} \mathbf{P}\left(\bigcap_{k=1}^{n} E_k\right),$$

for any finite sequence $(E_k)_{k=1}^{n}$ *in* $\mathcal{E}$.

Proof. . □

4.3.3 Countably Additive Probabilities

Let Ω be a sample space, and let $\mathcal{E}$ be a σ-algebra of events of Ω.

Definition 278 *We say that a function* $\mathbf{P} : \mathcal{E} \to \mathbb{R}_+$ *is a* countably additive probability *on* Ω *if it has the following properties*

1. $\mathbf{P}(\Omega) = 1$;
2. $\mathbf{P}\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{n=1}^{\infty} \mathbf{P}\left(E_n\right)$ *for every sequence* $(E_n)_{n\geq 1}$ *of pairwise incompatible events in* $\mathcal{E}$.

Definition 279 *If* $\mathbf{P} : \mathcal{E} \to \mathbb{R}_+$ *is a countably additive probability, Then, for every* $E \in \mathcal{E}$ *the positive real number* $\mathbf{P}\left(E\right)$ *is still called the* probability *of the event* E.

Example 280 (Dirac probability) *The Dirac probability (see Example 262 and Definition 263) is a countably additive probability.*

Discussion. Note first that the definition of the function $\mathbf{P}_0 : \mathcal{E} \to \mathbb{R}_+$ in Example 262 also applies in case $\mathcal{E}$ is a σ-algebra of events of Ω. Therefore, to show the claim, we just need to prove that 2 holds true. In fact, we have

$$\mathbf{P}_0\left(\bigcup_{n=1}^{\infty} E_n\right) = \begin{cases} 1, & \text{if } \omega_0 \in \bigcup_{n=1}^{\infty} E_n, \\ 0, & \text{if } \omega_0 \notin \bigcup_{n=1}^{\infty} E_n. \end{cases}$$

On the other hand, in case $\omega \in \bigcup_{n=1}^{\infty} E_n$, since the events in the sequence $(E_n)_{n\geq 1}$ are pairwise incompatible, there exists a unique n_0 such that $\omega_0 \in E_{n_0}$. Hence, $\mathbf{P}_0\left(E_{n_0}\right) = 1$ and $\mathbf{P}_0\left(E_n\right) = 0$, for every $n \neq n_0$. It follows

$$\sum_{n=1}^{\infty} \mathbf{P}\left(E_n\right) = 1.$$

Instead, in case $\omega_0 \notin \bigcup_{n=1}^{\infty} E_n$, clearly $\omega_0 \notin E_n$, for every $n \geq 1$. Hence, $\mathbf{P}_0\left(E_n\right) = 0$, for every $n \geq 1$. It follows

$$\sum_{n=1}^{\infty} \mathbf{P}\left(E_n\right) = 0.$$

Thus, the desired 2 holds true. □

Let $\mathbf{P} : \mathcal{E} \to \mathbb{R}_+$ be a countably additive probability on Ω.

Proposition 281 *We have*

1. $\mathbf{P}$ *is an additive probability;*
2. $\mathbf{P}\left(\bigcup_{n=1}^{\infty} E_n\right) \leq \sum_{n=1}^{\infty} \mathbf{P}\left(E_n\right)$ *for every sequence* $\left(E_n\right)_{n\geq 1}$ *in* $\mathcal{E}$.

Proof. With reference to 1, consider any pair E, F of incompatible events in $\mathcal{E}$. Then, the sequence $\left(E_n\right)_{n\geq 1}$ given by

$$E_1 \equiv E, \qquad E_2 \equiv F, \qquad E_n \equiv \varnothing, \quad \forall n \geq 3,$$

is a sequence of pairwise incompatible events in $\mathcal{E}$ such that

$$\bigcup_{n=1}^{\infty} E_n = E \cup F.$$

Since $\mathbf{P} : \mathcal{E} \to \mathbb{R}_+$ is a countably additive probability on Ω, we have

$$\mathbf{P}\left(E \cup F\right) = \mathbf{P}\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{n=1}^{\infty} \mathbf{P}\left(E_n\right) = \sum_{n=1}^{2} \mathbf{P}\left(E_n\right) = \mathbf{P}\left(E\right) + \mathbf{P}\left(F\right),$$

which proves that $\mathbf{P} : \mathcal{E} \to \mathbb{R}_+$ is also finitely additive. □

Note that, likewise Property 2 in Definition 230, Property 2 in Definition 278 is necessary to exploit the techniques of the infinitesimal calculus. For instance, we have

Proposition 282 *Let* $\left(E_n\right)_{n\geq 1}$ *be a sequence of events in* $\mathcal{E}$. *Assume* $\left(E_n\right)_{n\geq 1}$ *is increasing [risp. decreasing], that is*

$$E_n \subseteq E_{n+1} \qquad \text{[resp. } E_n \supseteq E_{n+1}\text{]}, \tag{4.42}$$

for every $n \in \mathbb{N}$. *Then,*

$$\mathbf{P}\left(\bigcup_{n=1}^{\infty} E_n\right) = \lim_{n\to\infty} \mathbf{P}(E_n) \qquad \text{[resp. } \mathbf{P}\left(\bigcap_{n=1}^{\infty} E_n\right) = \lim_{n\to\infty} \mathbf{P}(E_n)\text{]}. \tag{4.43}$$

Proof. Write

$$E_0 \equiv \varnothing, \quad F_n \equiv E_n - E_{n-1}, \quad \forall n \in \mathbb{N}.$$

Since the sequence $\left(E_n\right)_{n\geq 1}$ is increasing, the events of the sequence $\left(F_n\right)_{n\geq 1}$ are pairwise incompatible and we have

$$\bigcup_{n=1}^{\infty} E_n = \bigcup_{n=1}^{\infty} F_n,$$

In fact, let $n_1, n_2 \in \mathbb{N}$ such that $n_1 \neq n_2$, say $n_1 < n_2$. Then, $F_{n_1} \equiv E_{n_1} - E_{n_1-1} \subseteq E_{n_1}$ and $F_{n_2} \equiv E_{n_2} - E_{n_2-1}$ is incompatible with E_{n_2-1}. Since $E_{n_2-1} \supseteq E_{n_1}$, a fortiori F_{n_2} is incompatible with E_{n_1}. It clearly follows that F_{n_2} is also incompatible with F_{n_1}. Furthermore,

$$\begin{aligned}
\bigcup_{n=1}^{\infty} F_n &= \bigcup_{n=1}^{\infty} \left(E_n - E_{n-1}\right) = \bigcup_{n=1}^{\infty} \left(E_n \cap E_{n-1}^c\right) = \left(\bigcup_{n=1}^{\infty} E_n\right) \cap \left(\bigcup_{n=1}^{\infty} E_{n-1}^c\right) \\
&= \left(\bigcup_{n=1}^{\infty} E_n\right) \cap \left(\bigcap_{n=1}^{\infty} E_{n-1}\right)^c = \left(\bigcup_{n=1}^{\infty} E_n\right) \cap E_0^c = \left(\bigcup_{n=1}^{\infty} E_n\right) \cap \Omega \\
&= \bigcup_{n=1}^{\infty} E_n.
\end{aligned}$$

Recalling Property 3, we Then, obtain

$$\begin{aligned}
\mathbf{P}\left(\bigcup_{n=1}^{\infty} E_n\right) &= \mathbf{P}\left(\bigcup_{n=1}^{\infty} F_n\right) = \sum_{n=1}^{\infty} \mathbf{P}\left(F_n\right) = \lim_{n\to\infty} \sum_{k=1}^{n} \mathbf{P}\left(F_k\right) \\
&= \lim_{n\to\infty} \sum_{k=1}^{n} \mathbf{P}\left(E_n - E_{n-1}\right) = \lim_{n\to\infty} \sum_{k=1}^{n} \left(\mathbf{P}\left(E_n\right) - \mathbf{P}\left(E_{n-1}\right)\right) \\
&= \lim_{n\to\infty} \left(\mathbf{P}\left(E_n\right) - \mathbf{P}\left(E_0\right)\right) = \lim_{n\to\infty} \mathbf{P}\left(E_n\right).
\end{aligned}$$

To be continued. □

Remark 283 *Let $\mathcal{E}$ be a finite algebra of events of Ω. Then, any additive probability $\mathbf{P} : \mathcal{E} \to \mathbb{R}_+$ is a countably additive probability.*

From now on, we will speak of *probability* with the meaning of *countably additive probability*, retaining the term *finitely additive probability* in the sense of Definition 258.

4.3.4 Discrete Probability Densities: the finite case

Let $(p_k)_{k=1}^n$ be a finite sequence of positive real numbers.

Definition 284 *We say that $(p_k)_{k=1}^n$ is a* finite probability density *if*

$$\sum_{k=1}^{n} p_k = 1. \tag{4.44}$$

Let Ω be a sample space and let $\mathcal{E}$ be an algebra of events of Ω. Assume Ω is finite, say $\Omega \equiv \{\omega_1, \dots \omega_n\}$ for some $n \in \mathbb{N}$, and let $\mathbf{P}_k : \mathcal{P}(\Omega) \to \mathbb{R}_+$ be the Dirac probability concentrated at ω_k, for $k = 1, \dots, n$.

Proposition 285 (discrete probability) *If $(p_k)_{k=1}^n$ is a finite probability density, Then, the function $\mathbf{P} : \mathcal{E} \to \mathbb{R}_+$ given by*

$$\mathbf{P}(E) \stackrel{def}{=} \sum_{k=1}^{n} p_k \mathbf{P}_k(E), \quad \forall E \in \mathcal{E} \tag{4.45}$$

is the unique probability on Ω such that

$$\mathbf{P}(\omega_k) = p_k, \tag{4.46}$$

for every $n = 1, \dots, N$, where $\mathbf{P}(\omega_k)$ is the standard shorthand for $\mathbf{P}(\{\omega_k\})$.

Proof. To prove that $\mathbf{P}$ is a probability, we have to check that 1 and 2 of Definition 278 hold true. Since $\mathbf{P}_k$ is a probability, for $k = 1, \dots, n$, thanks to (4.44), we clearly have

$$\mathbf{P}(\Omega) = \sum_{k=1}^{n} p_k \mathbf{P}_k(\Omega) = \sum_{k=1}^{n} p_k = 1.$$

Now, since $\mathbf{P}_k$ is a probability, for $k = 1, \dots, n$, for any pair $E, F \in \mathcal{E}$ such that $E \cap F = \varnothing$, we have

$$\begin{aligned} \mathbf{P}(E \cup F) &= \sum_{k=1}^{n} p_k \mathbf{P}_k(E \cup F) = \sum_{k=1}^{n} p_k \left(\mathbf{P}_k(E) + \mathbf{P}_k(F)\right) \\ &= \sum_{k=1}^{n} p_k \mathbf{P}_k(E) + \sum_{k=1}^{n} p_k \mathbf{P}_k(F) \\ &= \mathbf{P}(E) + \mathbf{P}(F). \end{aligned}$$

This completes the proof that $\mathbf{P}$ is a probability. We are left with checking that $\mathbf{P}$ is the only probability which satisfies (4.46). We have

$$\mathbf{P}(\omega_k) = \sum_{j=1}^{n} p_j \mathbf{P}_j(\omega_k),$$

for every $k = 1, \ldots, n$. On the other hand, since $\mathbf{P}_j$ is the Dirac probability concentrated at ω_k, for $j = 1, \ldots, n$, we have

$$\mathbf{P}_j(\omega_k) = \begin{cases} 1 & \text{if } j = k \\ 0 & \text{if } j \neq k \end{cases},$$

for every $j = 1, \ldots, n$. Thus, Equation (4.46) immediately follows. In the end, assume that $\mathbf{Q} : \mathcal{P}(\Omega) \to \mathbb{R}$ is another probability fulfilling Equation (4.46). Then, we have

$$\mathbf{Q}(\omega_k) = \mathbf{P}(\omega_k).$$

for every $k = 1, \ldots, n$. It easily follows that

$$\mathbf{Q}(E) = \mathbf{P}(E)$$

for every $E \in \mathcal{P}(\Omega)$. In fact, we can clearly write

$$E = \bigcup_{\{k \in \{1,\ldots,n\} : \omega_k \in E\}} \{\omega_k\}.$$

As a consequence,

$$\begin{aligned} \mathbf{Q}(E) = \mathbf{Q}\left(\bigcup_{\{k \in \{1,\ldots,n\} : \omega_k \in E\}} \{\omega_k\} \right) &= \sum_{\{k \in \{1,\ldots,n\} : \omega_k \in E\}} \mathbf{Q}(\omega_k) \\ &= \sum_{\{k \in \{1,\ldots,n\} : \omega_k \in E\}} \mathbf{P}(\omega_k) = \mathbf{P}\left(\bigcup_{\{k \in \{1,\ldots,n\} : \omega_k \in E\}} \{\omega_k\} \right) \\ &= \mathbf{P}(E). \end{aligned}$$

This completes the proof. □

Definition 286 (discrete probability) *We call the probability presented in Proposition 285 the* probability with density $(p_k)_{k=1}^n$.

Remark 287 *We have*

$$\mathbf{P}(E) = \sum_{k=1}^{n} p_k 1_E(\omega_k),$$

where $1_E : \Omega \to \mathbb{R}$ is the indicator function of the event E, for every $E \in \mathcal{E}$.

Example 288 (Dirac density) *The trivial finite sequence $(p_k)_{k=1}^1$ given by*

$$p_1 \stackrel{def}{=} 1,$$

is a finite probability density.

Definition 289 (Dirac density) *We call the probability density presented in Example 288 the* Dirac density.

Remark 290 (Dirac density) *The probability associated to the Dirac density is the Dirac probability.*

Example 291 (Bernoulli density) *Choose any $p \in (0,1)$ and write $q \equiv 1-p$. Then, the sequence $(p_k)_{k=1}^2$ given by*

$$p_k \overset{def}{=} \begin{cases} p & \text{if } k=1 \\ q & \text{if } k=2 \end{cases}, \tag{4.47}$$

is a finite probability density.

Definition 292 (Bernoulli density) *We call the probability density presented in Example 291 the* Bernoulli density *with "success" probability p.*

Remark 293 (Bernoulli density) *The probability associated to the Bernoulli density is the Bernoulli probability.*

Example 294 (Bernoulli density) *Consider an urn containing $N \geq 2$ balls of which M are white and $N-M$ are black, for $1 \leq M < N$. Draw a ball from the urn. The probability p [resp. q] that the drawn ball is white [resp. black] is given by*

$$p = \frac{M}{N} \quad [\text{resp. } q = \frac{N-M}{N}]. \tag{4.48}$$

Discussion. Assume the balls in the urn are numbered from 1 to N, the white balls being numbered from 1 to M. We can Then, consider as a sample space "a ball is drawn fom the urn" the set $\Omega \equiv \{1,\ldots,N\}$ and as the event "a white ball is drawn fom the urn" the set $W \equiv \{1,\ldots,M\}$. Applying the naive probability model (see Definition **??**) we obtain

$$\mathbf{P}(W) = \frac{|W|}{|\Omega|} = \frac{M}{N} \equiv p,$$

On the other hand, writing B for the event "a black ball is drawn fom the urn", we clearly have

$$B = W^c,$$

which implies

$$\mathbf{P}(B) = 1 - \mathbf{P}(W) = 1 - p = \frac{N-M}{N}.$$

Note that, by a straightforward appproach,

$$\mathbf{P}(B) = \frac{|B|}{|\Omega|} = \frac{|W^c|}{|\Omega|} = \frac{N-M}{N}.$$

□

Example 295 (discrete uniform density) *The sequence $(p_k)_{k=1}^n$ given by*

$$p_k \overset{def}{=} \frac{1}{n}, \quad \forall k = 1,\ldots,n. \tag{4.49}$$

is a finite probability density.

Definition 296 (discrete uniform density) *We call the probability density presented in Example 295 the* finite uniform density.

Remark 297 (discrete uniform density) *The probability with finite uniform density is the naive probability.*

Example 298 (binomial density) *Let $n \in \mathbb{N}$, choose any $p \in (0,1)$ and write $q \equiv 1-p$. Then, the sequence $(p_{n,k})_{k=0}^{n}$ given by*

$$p_{n,k} \overset{def}{=} \binom{n}{k} p^{n-k} q^k, \quad \forall k = 0, 1, \ldots, n. \tag{4.50}$$

is a discrete probability density.

Discussion. . □

Definition 299 (binomial density) *We call the probability density presented in Example 298 the* binomial density *with parameters n and p. More specifically, the parameter n [resp. p] is called the* number of trials *and [*success probability*]. In this context, the redundant parameter q is called the* failure probability.

Definition 300 (binomial density) *The probability with binomial density is called the* binomial probability.

The binomial probability is used to model finite sequence of repetitions in the same conditions (trials) of a random phenomenon with only two distinguishable possible outcomes, one of which is considered a "success". The number $p_{n,k}$ gives the probability of obtaining k successes in n trials.

In Statistics, the binomial probability is used to model the *sampling with replacement* from a population made by distinguishable individuals of two different types.

Example 301 (binomial density) *Consider an urn containing $N \geq 2$ balls of which M are white and $N-M$ are black, for $1 \leq M < N$. Write*

$$p \equiv \frac{M}{N} \quad \textit{and} \quad q \equiv \frac{N-M}{N}.$$

Draw n balls from the urn with replacement *of the drawn ball in the urn. Then, the probability $p_{n,k}$ that the drawn sample of n balls contains k white balls (and $n-k$ black balls) is given by*

$$p_{n,k} = \binom{n}{k} p^k q^{n-k}, \tag{4.51}$$

for every $k = 0, 1, \ldots, n$.

Discussion. From Example 294, we know that p [rep. q] is the probability that the first drawn ball is white [resp. black]. After the draw the ball is put back into the urn. As a consequence, the composition of the urn before the second draw is identical to the composition before the first draw. Therefore, the probability that the second drawn ball is white [resp. black] is clearly equal to the probability that the first drawn ball is white [resp. black]. On

the other hand, to deal with both the first and second draw it may be convenient to think on two urns with the same composition represented by the sample spaces $\Omega_1 \equiv \{1, \ldots, N\}$ and $\Omega_2 \equiv \{1, \ldots, N\}$, numbering the white balls in the first and second urn from 1 to M. Hence the sets $E_1 \equiv \{1, \ldots, M\}$ [resp. $E_2 \equiv \{1, \ldots, M\}$] represents the event "a white ball is drawn from the first [resp. second] urn". Hence, the sample space "a ball is drawn fom the first urn and a ball is drawn from the second urn" is represented by the set $\Omega = \Omega_1 \times \Omega_2$ and the event "a white ball is drawn fom the first urn and a white ball is drawn from the second urn" is represented by the set $E = E_1 \times E_2$. Applying the naive probability model (see Definition **??**) we obtain

$$\mathbf{P}(E) = \frac{|E|}{|\Omega|} = \frac{|E_1 \times E_2|}{|\Omega_1 \times \Omega_2|} = \frac{|E_1|\,|E_1|}{|\Omega_1|\,|\Omega_2|} = \frac{M^2}{N^2} = p^2.$$

This is also the probability that, drawning from a single urn two times with replacement, the first and second drawn balls are both white. Iterating the argument, it is not difficult to realize that, drawning from the urn k times with replacement, the probability that all the balls from the first to the kth are white is given by

$$\mathbf{P}(E_1 \times \cdots \times E_k) = \frac{|E_1 \times \cdots \times E_k|}{|\Omega_1 \times \cdots \times \Omega_k|} = \frac{|E_1| \cdot \cdots \cdot |E_k|}{|\Omega_1| \cdot \cdots \cdot |\Omega_k|} = \frac{M^k}{N^k} = p^k$$

In turn, it is not difficult to realize that, drawning from the urn n times with replacement, the probability that all the balls from the first to the kth are white and all the balls from the $k+1$th to the nth are black is given by

$$\begin{aligned}
\mathbf{P}\left(E_1 \times \cdots \times E_k \times E_{k+1}^c \times \cdots \times E_n^c\right) &= \frac{\left|E_1 \times \cdots \times E_k \times E_{k+1}^c \times \cdots \times E_n^c\right|}{|\Omega_1 \times \cdots \times \Omega_n|} \\
&= \frac{|E_1| \cdot \cdots \cdot |E_k| \cdot \left|E_{k+1}^c\right| \cdot \cdots \cdot |E_n^c|}{|\Omega_1 \times \cdots \times \Omega_n|} \\
&= \frac{M^k (N-M)^{n-k}}{N^n} \\
&= \frac{M^k}{N^k} \frac{(N-M)^{n-k}}{N^{n-k}} \\
&= p^k q^{n-k}.
\end{aligned}$$

However, we are not interested in the order according to which the white balls are sampled in n draws. The white balls may appear in the first k draws or in any other way we can select k elements from the set $\{1, \ldots, n\}$. The selections of these k elements, say $n_1, \ldots, n_k$, represents the positions of the k white balls in the sample and, by virtue of the replacing mechanism, all these selections have the same probability of realization. The number of possible selections of k elements from the set $\{1, \ldots, n\}$ is given by the number $\binom{n}{k}$ of all k-combinations from a set of n elements. Therefore, writing $W_{n_1,\ldots,n_k}$ for the event "a white ball is drawn fom the urn at the n_1th, n_2th,..., n_kth draw", the probability that the drawn sample contains k white balls turns out to be

$$\begin{aligned}
\sum_{\{n_1,\ldots,n_k\} \subseteq \{1,\ldots,n\}} \mathbf{P}(W_{n_1,\ldots,n_k}) &= \sum_{\{n_1,\ldots,n_k\} \subseteq \{1,\ldots,n\}} \mathbf{P}(W_{1,\ldots,k}) \\
&= \sum_{\{n_1,\ldots,n_k\} \subseteq \{1,\ldots,n\}} \mathbf{P}\left(E_1 \times \cdots \times E_k \times E_{k+1}^c \times \cdots \times E_n^c\right) \\
&= \binom{n}{k} p^k q^{n-k},
\end{aligned}$$

as desired. □

Later, we will discuss again Example 301 by applying the notion of independent events.

Example 302 (multinomial density) *Let $m, n \in \mathbb{N}$ and let $p_1, \ldots, p_m \in (0,1)$ such that $\sum_{k=1}^{m} p_k = 1$. Set*

$$K_n^m \equiv \{(k_1, \ldots, k_m) \in \mathbb{N}_0^m : \textstyle\sum_{\ell=1}^m k_\ell = n\}. \tag{4.52}$$

Then, the sequence $(p_{n,k_1,\ldots,k_m})_{(k_1,\ldots,k_m)\in K_n^m}$ given by

$$p_{n,k_1,\ldots,k_m} \stackrel{def}{=} \frac{n!}{k_1! \cdots k_m!} p_1^{k_1} \cdot \cdots \cdot p_m^{k_m}, \quad \forall\, (k_1, \ldots, k_m) \in K_n^m, \tag{4.53}$$

is a finite probability density.

Discussion. . □

Definition 303 (multinomial density) *We call the probability density presented in Example 302 the* multinomial density *of parameters n and $p_1, \ldots, p_m$.*

In Statistics, the multinomial density is used to model the *sampling with replacement* from a population which contains distinguishable individuals of more than two different types.

Example 304 (multinomial density) *Consider an urn containing $N \geq 3$ balls of $m \geq 3$ different colors of which M_1 are white, M_2 are red, ..., and M_m are black, where $1 \leq M_\ell$, for every $\ell = 1, \ldots, m$, and $\sum_{\ell=1}^m M_\ell = N$. Write*

$$p_\ell \equiv \frac{M_\ell}{N}, \quad \forall \ell = 1, \ldots, m.$$

Draw n balls from the urn with replacement *of the drawn ball in the urn. Then, the probability $p_{n,k_1,\ldots,k_m}$ that the drawn sample of n balls contains k_1 white balls, k_2 red balls, ..., k_m black balls, where $k_\ell \geq 0$, for every $\ell = 1, \ldots, m$, and $\sum_{\ell=1}^m k_\ell = n$, is given by the multinomial density (4.53).*

Discussion. . □

Example 305 (bivariate hypergeometric density) *Let $M, N \in \mathbb{N}$ such that $M \leq N$. Let $n \in \mathbb{N}$ such that $n \leq \min\{M, N-M\}$. Then, the sequence $(p_{n,k})_{k=0}^n$ given by*

$$p_{n,k} \stackrel{def}{=} \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}, \quad \forall k = 0, 1, \ldots, n, \tag{4.54}$$

is a finite probability density.

Discussion. The positivity of $p_{n,k}$ being evident for any choice of $n \leq \min\{M, N-M\}$ and every $k = 0, 1, \ldots, n$, we just need to prove that (4.44) holds true, but this has altrady been done in the discussion of Example 273. □

Definition 306 (bivariate hypergeometric density) *We call the probability density introduced in Example 305 the* bivariate hypergeometric density *with parameters n and $p \equiv \frac{M}{N}$. More specificallly the parameter n [resp. p] is called the* number of trials *[*success probability*].*

In Statistics, the bivariate hypergeometric density is used to model the *sampling without replacement* from a population made by distinguishable individuals of two different types.

Example 307 (bivariate hypergeometric density) *Consider an urn containing $N \geq 2$ balls of which M are white and $N - M$ are black, for $1 \leq M < N$. Draw n balls from the urn* without replacement *of the drawn ball in the urn, where $n \leq \min\{M, N - M\}$. Then, the probability $p_{n,k}$ that the drawn sample of n balls contains $k \leq n$ white balls is given by the bivariate hypergeometric density (4.54).*

Discussion. Assume the balls in the urn are numbered from 1 to N, the white balls being numbered from 1 to M. Hence, drawing $n \leq N$ balls from the urn without replacement is the same than sampling a subset of n elements from the set $\{1, \ldots, N\}$. We can Then, assume that our sample space Ω is the family of all subsets of n elements from $\{1, \ldots, N\}$. We Then, have $|\Omega| = \binom{N}{n}$. In this setting, a favorable outcome turns out to be any subset in Ω containing k numbers chosen from $\{1, \ldots, M\}$ and $n-k$ numbers chosen from $\{M + 1, \ldots, N\}$. There are $\binom{M}{k}$ ways of chosing k numbers from $\{1, \ldots, M\}$ and for each of these choices there are $\binom{N-M}{n-k}$ ways of chosing $n - k$ numbers from $\{M + 1, \ldots, N\}$. Therefore, the number of favorable outcomes is given by $\binom{M}{k}\binom{N-M}{n-k}$. The desired claim immediately follows. □

Proposition 308 (bivariate hypergeometric density) *Assume the size N of the population is much larger that the size n of drawn sample. Then, the hypergeometric density of paramenters n and p, given by (4.54), converges to the binomial density of parameters n and p, given by (4.50).*

Proof. By virtue of the properties of the binomial coefficient, we have

$$\begin{aligned}
\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} &= \frac{\frac{M!}{k!(M-k)!}\frac{(N-M)!}{(n-k)!(N-M-(n-k))!}}{\frac{N!}{n!(N-n)!}} \\
&= \frac{M!}{k!\,(M-k)!}\frac{(N-M)!}{(n-k)!\,(N-M-(n-k))!}\frac{n!\,(N-n)!}{N!} \\
&= \frac{n!}{k!\,(n-k)!}\frac{M!}{(M-k)!}\frac{(N-M)!}{(N-M-(n-k))!}\frac{(N-n)!}{N!} \\
&= \binom{n}{k}\frac{(M-k+1)\cdot\cdots\cdot M\cdot(N-M-(n-k)+1)\cdot\cdots\cdot(N-M)}{(N-n+1)\cdot\cdots\cdot N} \qquad (4.55)
\end{aligned}$$

In addition, setting $p \equiv \frac{M}{N}$, we can rewrite

$$\begin{aligned}
&\frac{(M-k+1)\cdot\cdots\cdot M\cdot(N-M-(n-k)+1)\cdot\cdots\cdot(N-M)}{(N-n+1)\cdot\cdots\cdot N} \\
&= \frac{(pN-k+1)\cdot\cdots\cdot pN\cdot((1-p)N-(n-k)+1)\cdot\cdots\cdot(1-p)N}{(N-n+1)\cdot\cdots\cdot N}. \qquad (4.56)
\end{aligned}$$

Now, as N is very large with respect to n, we have

$$\begin{aligned}
&\underbrace{(pN-k+1)\cdot\dots\cdot pN}_{k \text{ factors}} \simeq p^k N^k,\\
&\underbrace{((1-p)N-(n-k)+1)\cdot\dots\cdot(1-p)N}_{n-k \text{ factors}} \simeq (1-p)^{n-k} N^{n-k}\\
&\underbrace{(N-n+1)\cdot\dots\cdot N}_{n \text{ factors}} \simeq N^n
\end{aligned} \tag{4.57}$$

Combining (4.55)-(4.57), it clearly follows

$$\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} \simeq \binom{n}{k} p^k (1-p)^{n-k},$$

as desired. □

Example 309 (multivariate hypergeometric density) *Let $M_1,\dots,M_m,N\in\mathbb{N}$ such that $1\le M_\ell$, for every $\ell=1,\dots,m$, and $\sum_{\ell=1}^m M_\ell = N$. Let $n\le \min\{M_1,\dots,M_m\}$ and set*

$$\tilde{K}_n^m \equiv \{(k_1,\dots,k_m)\in\mathbb{N}_0^m : k_\ell\le M_\ell,\ \forall \ell=1,\dots,m\ \wedge\ \textstyle\sum_{\ell=1}^m k_\ell = n\}. \tag{4.58}$$

Then, the sequence $(p_{n,k_1,\dots,k_m})_{(k_1,\dots,k_m)\in\tilde{K}_n^m}$ given by

$$p_{n,k_1,\dots,k_m} \overset{def}{=} \frac{\binom{M_1}{k_1}\binom{M_2}{k_2}\dots\binom{M_m}{k_m}}{\binom{N}{n}}, \quad \forall (k_1,\dots,k_m)\in\tilde{K}_n^m, \tag{4.59}$$

is a finite probability density.

Discussion. Assume to have a set Ω containing N elements, which can be partitioned, according to some characteristic of its elements, in m different subsets each Ω_k of which contains M_ℓ elements, where $1\le M_\ell$, for every $\ell=1,\dots,m$, and $\sum_{\ell=1}^m M_\ell = N$. For instance, consider an urn containing N balls of different colors of which M_1 are white, M_2 are red, ..., and M_m are black. Now, for any subset E of Ω we have

$$E = \textstyle\bigcup_{\ell=1}^m (E\cap\Omega_\ell).$$

Hence, any subset E of Ω containing n elements can be formed by picking $k_1 = |E\cap\Omega_1|$ elements from Ω_1, $k_2 = |E\cap\Omega_2|$ elements from Ω_2, ..., $k_m = |E\cap\Omega_m|$ elements from Ω_m, where $k_\ell\le M_\ell$, for every $\ell=1,\dots,m$, and $\sum_{\ell=1}^m k_\ell = n$. In turn, by picking k_1 elements from Ω_1 we can form $\binom{M_1}{k_1}$ distinct subsets of Ω containing k_1 elements. For each of these subsets, by picking k_2 elements from Ω_2 we can form $\binom{M_2}{k_2}$ distinct subsets of Ω containing k_1+k_2 elements. Therefore, by picking k_1 elements from Ω_1 and k_2 elements from Ω_2 we can form $\binom{M_1}{k_1}\binom{M_2}{k_2}$ distinct subsets of Ω containing k_1+k_2 elements. Iterating this argument, it is not difficult to realize that the number

$$\binom{M_1}{k_1}\binom{M_2}{k_2}\dots\binom{M_m}{k_m}$$

represents the number of all subsets of Ω containing $\sum_{\ell=1}^{m} k_\ell = n$ elements that can be formed by picking m_ℓ elements from Ω_ℓ, for every $\ell = 1, \ldots, m$. On the other hand, on varying the number k_ℓ of elements from Ω_k, for every $\ell = 1, \ldots, m$, we can form any subsets of Ω containing n elements. This shows that the generalized Vandermonde identity

$$\sum_{(k_1,\ldots,k_\ell)\in\tilde{K}_n^m} \binom{M_1}{k_1}\binom{M_2}{k_2}\cdots\binom{M_m}{k_m} = \binom{N}{n},$$

holds true. As a consequence, we have

$$\sum_{(k_1,\ldots,k_\ell)\in\tilde{K}_n^m} p_{n,k_1,\ldots,k_\ell} = \sum_{(k_1,\ldots,k_\ell)\in\tilde{K}_n^m} \frac{\binom{M_1}{k_1}\binom{M_2}{k_2}\cdots\binom{M_m}{k_m}}{\binom{N}{n}} = 1,$$

which shows that the finite sequence given by (4.59) is a probability density. □

Definition 310 (multivariate hypergeometric density) *We call the probability density introduced in Example 309 the* multivariate hypergeometric density *of parameters N and $p_1 \equiv \frac{M_1}{N}, \ldots, p_m \equiv \frac{M_m}{N}$.*

In Statistics, the multivariate hypergeometric density is used to model the *sampling without replacement* from a population made by distinguishable individuals of more than two different types..

Example 311 *Consider an urn containing $N \geq 3$ balls of $m \geq 3$ different colors of which M_1 are white, M_2 are red, ..., and M_m are black, where $1 \leq M_\ell$, for every $\ell = 1, \ldots, m$, and $\sum_{\ell=1}^{m} M_\ell = N$. Draw $n \leq \min\{M_1, \ldots, M_m\}$ balls from the urn* without replacement *of the drawn ball in the urn. Then, the probability $p_{n,k_1,\ldots,k_m}$ that the drawn sample of n balls contains k_1 white balls, k_2 red balls, ..., k_m black balls, where $k_\ell \geq 0$, for every $\ell = 1, \ldots, m$, and $\sum_{\ell=1}^{m} k_\ell = n$, is given by the multivariate hypergeometric density (4.59).*

Discussion. . □

4.3.5 Discrete Probability Densities: the denumerable case

Let $(p_n)_{n\geq 1}$ be a sequence of real positive numbers.

Definition 312 *We say that* $(p_n)_{n\geq 1}$ *is a* denumerable probability density *if*

$$\sum_{n=1}^{\infty} p_n = 1. \tag{4.60}$$

Let Ω be a sample space and let $\mathcal{E}$ be a σ-algebra of events of Ω. Assume Ω is denumerable, that is $\Omega \equiv \{\omega_n : n \in \mathbb{N}\}$ and let $\mathbf{P}_n : \mathcal{E} \to \mathbb{R}_+$ be the Dirac probability concentrated at ω_n, for any $n \in \mathbb{N}$.

Proposition 313 *If* $(p_n)_{n\geq 1}$ *is a denumerable probability density, Then, the function* $\mathbf{P} : \mathcal{E} \to \mathbb{R}$ *given by*

$$\mathbf{P}(E) \stackrel{def}{=} \sum_{n=1}^{\infty} p_n \mathbf{P}_n(E), \quad \forall E \in \mathcal{E} \tag{4.61}$$

is the unique probability on Ω *such that*

$$\mathbf{P}(\omega_n) = p_n, \tag{4.62}$$

for every $n \in \mathbb{N}$*, where* $\mathbf{P}(\omega_n)$ *is the standard abbreviation for* $\mathbf{P}(\{\omega_n\})$*.*

Proof. Since $\mathbf{P}_n : \mathcal{E} \to \mathbb{R}_+$ is a countably additive probability (see Example 280) and Equation 4.60 holds true, applying Equation (4.61) we obtain

$$\mathbf{P}(\Omega) = \sum_{n=1}^{\infty} p_n \mathbf{P}_n(\Omega) = \sum_{n=1}^{\infty} p_n = 1,$$

which shows that Equation (1) holds true. In addition,

$$\mathbf{P}\left(\bigcup_{m=1}^{\infty} E_m\right) = \sum_{n=1}^{\infty} p_n \mathbf{P}_n\left(\bigcup_{m=1}^{\infty} E_m\right) = \sum_{n=1}^{\infty} p_n \left(\sum_{m=1}^{\infty} \mathbf{P}_n(E_m)\right) = \sum_{n=1}^{\infty}\left(\sum_{m=1}^{\infty} p_n \mathbf{P}_n(E_m)\right) \tag{4.63}$$

for every sequence $(E_m)_{m\geq 1}$ of pairwise incompatible events in $\mathcal{E}$. Now,

$$\sum_{m=1}^{\infty} p_n \mathbf{P}_n(E_m) \leq p_n$$

for every $n \in \mathbb{N}$. In fact, since the events of the sequence $(E_m)_{m\geq 1}$ are mutually incompatible there exists at most one $m_n \in \mathbb{N}$ such that $\mathbf{P}_n(E_{m_n}) = 1$ and $\mathbf{P}_n(E_m) = 0$ for every $m \neq m_n$. As a consequence, we have

$$\sum_{n=1}^{\infty}\left(\sum_{m=1}^{\infty} p_n \mathbf{P}_n(E_m)\right) \leq \sum_{n=1}^{\infty} p_n = 1.$$

This implies that we can write

$$\sum_{n=1}^{\infty}\left(\sum_{m=1}^{\infty} p_n \mathbf{P}_n(E_m)\right) = \sum_{m=1}^{\infty}\left(\sum_{n=1}^{\infty} p_n \mathbf{P}_n(E_m)\right) = \sum_{m=1}^{\infty} \mathbf{P}(E_m). \tag{4.64}$$

Combining (4.63) and (4.64), it follows that also Equation (2) holds true. In the end, to show that $\mathbf{P} : \mathcal{E} \to \mathbb{R}$ is the unique probability satisfying (4.62), we can just repeat the argument in the last part of the proof of Proposition 285. □

Definition 314 (probability associated to a denumerable density) *We call the probability introduced in Proposition 313 the* probability with density $(p_n)_{n\geq 1}$.

Remark 315 *We have*

$$\mathbf{P}(E) = \sum_{n=1}^{\infty} p_n 1_E(\omega_n),$$

where $1_E : \Omega \to \mathbb{R}$ is the indicator function of the event E, for every $E \in \mathcal{E}$.

Example 316 (geometric density) *Choose any $p \in (0,1)$ and write $q \equiv 1-p$. Then, the sequence $(p_n)_{n\geq 0}$ given either by*

$$p_n \overset{def}{=} pq^{n-1}, \quad \forall n \geq 1. \tag{4.65}$$

or

$$p_n \overset{def}{=} pq^{n}, \quad \forall n \geq 0. \tag{4.66}$$

is a denumerable probability density.

Discussion. .The sequence $(p_n)_{n\geq 0}$ characterized by either (4.65) or (4.66) is clearly positive. Hence, we need only to show that (4.60) holds true. In fact, we have

$$\sum_{n=1}^{\infty} pq^{n-1} = \sum_{n=0}^{\infty} pq^{n} = p\sum_{n=0}^{\infty} q^{n} = \frac{p}{1-q} = \frac{p}{p} = 1,$$

as desired. □

Definition 317 (geometric density) *We call both the two sequences $(p_n)_{n\geq 0}$ introduced in Example 316 the* geometric density *with* success probability *p. The probability associated to the geometric density will be called the* geometric probability.

The geometric density is used to model a sequence of identical repetitions of any random phenomenon with only two distinguishable possible outcomes, of which one is considered a "success" and the other is considered a "failure". With reference to (4.65) [resp. (4.66), the number p_n gives the probability of obtaining the first success at the nth trial of the experiment [resp. the probability of obtaining the first success after n failures].

Example 318 (geometric density) *A coin is flipped repeatedly. Write p [resp. q] for the probability of getting "heads" [resp. "tails"] at the single flip. Then, the probability of getting "heads" for the first time at an odd [resp. even] flip turns out to be $\frac{1}{1+q}$ [resp. $\frac{q}{1+q}$].*

Discussion. Consider the following sequences of flips

$$\begin{aligned}
\omega_1 &\equiv (h, ?, ?, ?, \dots, ?, \dots), \\
\omega_2 &\equiv (t, h, ?, ?, \dots, ?, \dots), \\
\omega_3 &\equiv (t, t, h, ?, \dots, ?, \dots), \\
&\dots \\
\omega_n &\equiv (t, t, t, t, \dots, \underset{n}{h}, ? \dots), \\
&\dots
\end{aligned}$$

where the question mark denotes that we are not interested in the further outcomes of the flips. Introduce the sample space $\Omega \equiv \{\omega_n\}_{n\geq 1}$ and consider the probability associated to the density density $(p_n)_{n\geq 1}$. The elementary event $\{\omega_n\}$ models the event "the first heads occurs at the nth flip". We have

$$\mathbf{P}(\omega_n) = pq^{n-1}, \quad \forall n \geq 1. \tag{4.67}$$

To realize this, recall that in Example 301 is discussed the probability of getting an ordered sequence of k initial white balls and $n-k$ final black balls drawing n balls with replacement from an urn with M white balls and $N-M$ black balls. Essentially the same argument yields (4.67). In addition, the events $\{\omega_{n_1}\}$ and $\{\omega_{n_2}\}$ are mutually incompatible for all $n_1, n_2 \geq 1$ such that $n_1 \neq n_2$. Now, write $E_{\mathbb{O}}$ for the event "the first heads occurs at an odd flip". We have clearly

$$E_{\mathbb{O}} = \bigcup_{n=1}^{\infty} \{\omega_{2n-1}\}.$$

Hence,

$$\mathbf{P}(E_{\mathbb{O}}) = \mathbf{P}\left(\bigcup_{n=1}^{\infty} \{\omega_{2n-1}\}\right) = \sum_{n=1}^{\infty} \mathbf{P}(\omega_{2n-1}) = \sum_{n=1}^{\infty} pq^{2n-2} = p\sum_{n=1}^{\infty} \left(q^2\right)^{n-1} = \frac{p}{1-q^2} = \frac{1}{1+q}.$$

Similarly, writing $E_{\mathbb{E}}$ for the event "the first heads occurs at an even flip". We have

$$E_{\mathbb{E}} = \bigcup_{n=1}^{\infty} \{\omega_{2n}\}.$$

Thus,

$$\mathbf{P}(E_{\mathbb{E}}) = \mathbf{P}\left(\bigcup_{n=1}^{\infty} \{\omega_{2n}\}\right) = \sum_{n=1}^{\infty} \mathbf{P}(\omega_{2n}) = \sum_{n=1}^{\infty} pq^{2n-1} = \frac{p}{q}\sum_{n=1}^{\infty} \left(q^2\right)^{n} = \frac{p}{q}\frac{q^2}{1-q^2} = \frac{q}{1+q}.$$

Note that

$$\mathbf{P}(E_{\mathbb{O}}) + \mathbf{P}(E_{\mathbb{E}}) = 1.$$

Indeed, the events $E_{\mathbb{O}}$ and $E_{\mathbb{E}}$ constitute a partition of Ω. □

Example 319 (Poisson density) *Let $\lambda > 0$, and set*

$$p_n \overset{def}{=} e^{-\lambda}\frac{\lambda^n}{n!}, \quad \forall n \geq 0. \tag{4.68}$$

Then, the sequence $(p_n)_{n\geq 0}$ is a probability density on Ω.

Discussion. The sequence $(p_n)_{n\geq 0}$ being clearly positive, we need only to show that Condition 4.60 holds true. In fact, we have

$$\sum_{n=0}^{\infty} p_n = \sum_{n=0}^{\infty} e^{-\lambda}\frac{\lambda^n}{n!} = e^{-\lambda}\sum_{n=0}^{\infty}\frac{\lambda^n}{n!} = e^{-\lambda}e^{\lambda} = 1,$$

as desired. □

Definition 320 (Poisson density) *We call the sequence defined by (4.68) the* Poisson density *with* rate *or* intensity *parameter* λ. *The probability with Poisson density will be called the* Poisson probability.

The Poisson probability is applied to model the number of times in which an outcome of a random phenomenon may occur in a given unit of time. The parameter λ is the average number of times in which the considered outcome occurs in the given unit of time. The number p_n gives the probability that the outcome of the random phenomenon occurs exactly n times in the given unit of time.

Example 321 (Poisson density) *Assume over the period of one year, the average number of messages arriving per day at a smartphone is* 4. *Consider the probability that at some day the owner of the smartphone gets* 0, 1, 2, 3, 4, 5, 6 *messages, not more than* 3 *messages, not less than* 5 *messages, an odd [resp. even] number of messages.*

Discussion. Since the average number of messages arriving per day at the smartphone is supposed to be 4, we set $\lambda \equiv 4$. The probability of getting 0, 1, 2, 3, 4, 5, 6 messages is Then, given by

$$\begin{aligned}
p_0 &= e^{-4}\frac{\lambda^0}{0!} = e^{-4} = 1.8316\times 10^{-2},\\
p_1 &= e^{-4}\frac{4^1}{1!} = 4e^{-4} = 7.3263\times 10^{-2},\\
p_2 &= e^{-4}\frac{4^2}{2!} = 8e^{-4} = 0.14653,\\
p_3 &= e^{-4}\frac{4^3}{3!} = \frac{32}{3}e^{-4} = 0.19537,\\
p_4 &= e^{-4}\frac{4^4}{4!} = \frac{32}{3}e^{-4} = 0.19537,\\
p_5 &= e^{-4}\frac{4^5}{5!} = \frac{128}{15}e^{-4} = 9.539\,4\times 10^{-2},\\
p_6 &= e^{-4}\frac{4^6}{6!} = \frac{256}{45}e^{-4} = 3.296\,8\times 10^{-2}
\end{aligned}$$

Note that on varying of $n \geq 0$ the probability p_n initially increases, it attains the maximum at $n = 3$ and $n = 4$, thereafter p_n monotonically decreases to 0. Now, writing E_n for the elementary event "the owner of the smartphone gets n messages", the events "the owner of the smartphone gets not more than 4 messages" and "the owner of the smartphone gets not less than 4 messages" are clearly representable as $\bigcup_{n\leq 4} E_n$ and $\bigcup_{n\geq 4} E_n$, respectively. Therefore,

since the events E_{n_1} and E_{n_2} are pairwise incompatible for all $n_1, n_2 \geq 1$ such that $n_1 \neq n_2$, we have

$$\begin{aligned}\mathbf{P}\left(\bigcup_{n\leq 4} E_n\right) &= \sum_{n=0}^{4} \mathbf{P}\left(E_n\right) = e^{-4}\sum_{n=0}^{4}\frac{4^n}{n!} \\ &= \left(1+4+8+\frac{32}{3}+\frac{32}{3}\right)e^{-4} = \frac{103}{3}e^{-4} = 0.62884,\end{aligned}$$

and

$$\begin{aligned}\mathbf{P}\left(\bigcup_{n\geq 4} E_n\right) &= \sum_{n=4}^{\infty} \mathbf{P}\left(E_n\right) = \sum_{n=0}^{\infty} \mathbf{P}\left(E_n\right) - \sum_{n=0}^{3} \mathbf{P}\left(E_n\right) = 1 - e^{-4}\sum_{n=0}^{3}\frac{4^n}{n!} \\ &= 1 - \left(1+4+8+\frac{32}{3}\right)e^{-4} = 1 - \frac{71}{3}e^{-4} = 0.56653.\end{aligned}$$

In terms of the elementary events E_n, the event "the owner of the smartphone gets an odd [resp. even] number of messages" can be represented by $\bigcup_{n\in\mathbb{O}} E_n$ [resp. $\bigcup_{n\in\mathbb{E}} E_n$]. We Then, have

$$\begin{aligned}\mathbf{P}\left(\bigcup_{n\in\mathbb{O}} E_n\right) &= \mathbf{P}\left(\bigcup_{n\geq 0} E_{2n+1}\right) = \sum_{n=0}^{\infty} \mathbf{P}\left(E_{2n+1}\right) \\ &= e^{-4}\sum_{n=0}^{\infty}\frac{4^{2n+1}}{(2n+1)!} = e^{-4}\left(\frac{e^4 - e^{-4}}{2}\right) \\ &= \frac{1-e^{-8}}{2}.\end{aligned}$$

and

$$\begin{aligned}\mathbf{P}\left(\bigcup_{n\in\mathbb{E}} E_n\right) &= \mathbf{P}\left(\bigcup_{n\geq 0} E_{2n}\right) = \sum_{n=0}^{\infty} \mathbf{P}\left(E_{2n}\right) \\ &= e^{-4}\sum_{n=0}^{\infty}\frac{4^{2n}}{(2n)!} = e^{-4}\left(\frac{e^4 + e^{-4}}{2}\right) \\ &= \frac{1+e^{-8}}{2}.\end{aligned}$$

Note that the probability of getting an even number of messages is slightly higher than the probability of getting an odd number of messages. This is not surprising, though. Indeed, the average numbers of messages per day is even. □

4.3.6 Probability Densities on $\mathbb{R}^N$.

Let $\mathbb{R}^N$ be the real N-dimensional Euclidean space, for some $N \in \mathbb{N}$, and let $\mathcal{B}\left(\mathbb{R}^N\right)$ be the Borel σ-algebra of $\mathbb{R}^N$.

Definition 322 *A probability* $\mathbf{P} : \mathcal{B}\left(\mathbb{R}^N\right) \to \mathbb{R}_+$ *is also commonly referred to as a* probability distribution *on* $\mathbb{R}^N$.

The main difficulty in defining non-trivial probability distributions on $\mathbb{R}^N$ is represented by the vastity of the Borel σ-algebra $\mathcal{B}\left(\mathbb{R}^N\right)$. Nevertheless, this difficulty can be circumvented if we equip the measurable space $\left(\mathbb{R}^N, \mathcal{B}\left(\mathbb{R}^N\right)\right) \equiv \mathbb{R}^N$ with the Borel-Lebesgue measure $\mu_L^N : \mathcal{B}\left(\mathbb{R}^N\right) \to \mathbb{R}_+$. Thanks to this, a probability distribution can be easily assigned by means of a class of positive Borel functions, the so called *density functions*.

Let $f : \mathbb{R}^N \to \mathbb{R}$, briefly f, be a real function on $\mathbb{R}^N$.

Definition 323 *We say that* f *is a* Borel *function if the inverse images of Borel sets in* $\mathbb{R}$ *are Borel sets in* $\mathbb{R}^N$. *In symbols,*

$$f^{-1}(B) \in \mathcal{B}\left(\mathbb{R}^N\right) \quad \forall B \in \mathcal{B}(\mathbb{R}). \tag{4.69}$$

Let $f : \mathbb{R}^N \to \mathbb{R}$, briefly f, be a Borel real function on $\mathbb{R}^N$.

Definition 324 *We say that* f *is* almost everywhere positive *on* $\mathbb{R}^N$ *and we write* $f(x) \geq 0$ *a.e. on* $\mathbb{R}^N$ *if the inverse image of strictly negative half line has Borel-Lebesgue measure equal to zero. In symbols,*

$$\mu_L^N\left(f^{-1}(\mathbb{R}_{--})\right) = 0. \tag{4.70}$$

Let $f : \mathbb{R}^N \to \mathbb{R}$, briefly f, be a Borel real function on $\mathbb{R}^N$ which is almost everywhere positive on $\mathbb{R}^N$.

Definition 325 *We say that* f *is a* probability density *on* $\mathbb{R}^N$ *if we have*

$$\int_{\mathbb{R}^N} f(x)\, d\mu_L^N(x) = 1, \tag{4.71}$$

where $\int_{\mathbb{R}^N} f(x)\, d\mu_L^N(x)$ *is the Lebesgue integral of* f *on* $\mathbb{R}^N$.

Let $f : \mathbb{R}^N \to \mathbb{R}$ be a probability density on $\mathbb{R}^N$.

Theorem 326 *The function* $\mathbf{P}_f : \mathcal{B}\left(\mathbb{R}^N\right) \to \mathbb{R}_+$ *given by*

$$\mathbf{P}_f(B) \overset{def}{=} \int_B f(x)\, d\mu_L^N(x), \quad \forall B \in \mathcal{B}\left(\mathbb{R}^N\right), \tag{4.72}$$

is a probability distribution on $\mathbb{R}^N$, *called the* probability distribution with density f.

Proof. Recall that the non-negativity almost everywhere of f implies that the Lebesgue integral $\int_B f(x)\, d\mu_L^N(x)$ is non-negative for every $B \in \mathcal{B}(\mathbb{R}^N)$. Moreover, by virtue of Equation (4.71), we have

$$P_f(\mathbb{R}^N) = \int_{\mathbb{R}^N} f(x)\, d\mu_L^N(x) = 1.$$

In the end, thanks to the properties of the Lebesgue integral, we have

$$\mathbf{P}_f\left(\bigcup_{n=1}^{\infty} B_n\right) = \int_{\bigcup_{n=1}^{\infty} B_n} f(x)\, d\mu_L^N(x) = \sum_{n=1}^{\infty} \int_{B_n} f(x)\, d\mu_L^N(x) = \sum_{n=1}^{\infty} P_f(B_n),$$

for every sequence $(B_n)_{n\geq 1}$ of pairwise incompatible sets in $\mathcal{B}(\mathbb{R}^N)$. □

Proposition 327 *Given any probability density $f : \mathbb{R}^N \to \mathbb{R}$, we have*

$$\mathbf{P}_f(N) = 0$$

or every $B \in \mathcal{B}(\mathbb{R}^N)$ such that $\mu_L^N(B) = 0$. In particular, we have

$$\mathbf{P}_f(x) = 0,$$

for every $x \in \mathbb{R}^N$, where $\mathbf{P}_f(x)$ is the standard abbreviation for $\mathbf{P}_f(\{x\})$.

Proof. We have

$$\mathbf{P}_f(N) = \int_N f(x)\, d\mu_L^N(x) = \int_{\mathbb{R}^N} f(x)\, 1_N(x)\, d\mu_L^N(u).$$

Now, the function $g : \mathbb{R}^N \to \mathbb{R}$, given by

$$g(x) \overset{\text{def}}{=} f(x)\, 1_N(x), \quad \forall x \in \mathbb{R}^N$$

takes the value 0 almost everywhere on $\mathbb{R}^N$. Therefore, by virtue of the properties of the Lebesgue integral, we know that

$$\int_{\mathbb{R}^N} g(x)\, d\mu_L^N(u) = 0,$$

which proves the desired claim. □

Example 328 (continuous uniform density on $\mathbb{R}$) *Assume $N = 1$, fix any $a, b \in \mathbb{R}$ such that $a < b$, and consider the function $f_{a,b} : \mathbb{R} \to \mathbb{R}$ given by*

$$f_{a,b}(x) \overset{def}{=} \frac{1}{b-a} 1_{[a,b]}(x), \quad \forall x \in \mathbb{R}, \tag{4.73}$$

where $1_{[a,b]} : \mathbb{R} \to \mathbb{R}$ is the indicator function of the interval $[a, b]$. Then, $f_{a,b}$ is a probability density on $\mathbb{R}$.

Discussion. Considering the Borel-Lebesgue measure $\mu : \mathcal{B}(\mathbb{R}) \to \mathbb{R}_+$, by virtue of the properties of the Lebesgue integral, we have

$$\begin{aligned}\int_{\mathbb{R}} f_{a,b}(x)\, d\mu_L(x) &= \int_{\mathbb{R}} \frac{1}{b-a} 1_{[a,b]}(x) d\mu_L(x) = \frac{1}{b-a} \int_{\mathbb{R}} 1_{[a,b]}(x) d\mu_L(x) \\ &= \frac{1}{b-a} \mu([a,b]) = \frac{b-a}{b-a} = 1,\end{aligned}$$

as desired. □

Definition 329 (continuous uniform density on $\mathbb{R}$) *The probability density $f_{a,b} : \mathbb{R} \to \mathbb{R}$ defined by (4.73) is called the* continuous uniform density *on $\mathbb{R}$ with support the interval $[a, b]$. The probability distribution with continuous uniform density $f_{a,b}$ is called the* continuous uniform *probability distribution on $\mathbb{R}$ with support the interval $[a, b]$.*

Example 330 (continuous uniform density on $\mathbb{R}^N$) *For any $N \in \mathbb{N}$, fix any $a \equiv (a_1, \ldots, a_N)$, $b \equiv (b_1, \ldots, b_N) \in \mathbb{R}^N$ such that $a_n < b_n$, for every $n = 1, \ldots, N$, and consider the function $f_{a,b} : \mathbb{R}^N \to \mathbb{R}$ given by*

$$f_{a,b}(x) \overset{def}{=} \prod_{n=1}^{N} \frac{1}{b_k - a_k} 1_{[a,b]}(x), \quad \forall x \in \mathbb{R}^N, \tag{4.74}$$

where $1_{[a,b]} : \mathbb{R}^N \to \mathbb{R}$ is the indicator function of the interval $[a, b]$. Then, $f_{a,b}$ is a probability density on $\mathbb{R}^N$.

Definition 331 (continuous uniform density on $\mathbb{R}^N$) *The probability density $f_{a,b} : \mathbb{R}^N \to \mathbb{R}$ defined by (4.74) is called the* continuous uniform density *on $\mathbb{R}^N$ with support the interval $[a, b]$. The probability distribution with continuous uniform density $f_{a,b}$ is called the* continuous uniform *probability distribution on $\mathbb{R}^N$ with support the interval $[a, b]$.*

Example 332 (Cauchy density) *Assume $N = 1$ and let $f : \mathbb{R} \to \mathbb{R}$ be the function given by*

$$f(x) \overset{def}{=} \frac{1}{\pi} \frac{1}{1 + x^2}, \quad \forall x \in \mathbb{R}. \tag{4.75}$$

Then, f is a probability density on $\mathbb{R}$.

Discussion. First, since the function f is positive and continuous on $\mathbb{R}$, the Lebesgue integral $\int_{\mathbb{R}} f(x)\, d\mu_L(x)$ coincides with the improper Riemann integral $\int_{-\infty}^{+\infty} f(x)\, dx$. Second, a straightforward computation yields

$$\begin{aligned} \int_{-\infty}^{+\infty} \frac{1}{1 + x^2} dx &= \lim_{x \to +\infty} \int_{-x}^{x} \frac{1}{1 + u^2} du = \lim_{x \to +\infty} 2 \int_{0}^{x} \frac{1}{1 + u^2} dx \\ &= 2 \lim_{x \to +\infty} \arctan(u)|_0^x = 2 \lim_{x \to +\infty} (\arctan(x) - \arctan(0)) \\ &= \pi. \end{aligned}$$

It immediately follows that

$$\int_{\mathbb{R}} f(x)\, d\mu_L(x) = \int_{-\infty}^{+\infty} \frac{1}{\pi} \frac{1}{1 + x^2} dx = \frac{1}{\pi} \int_{-\infty}^{+\infty} \frac{1}{1 + x^2} dx = 1,$$

as desired. □

Definition 333 (Cauchy density) *The probability density $f : \mathbb{R} \to \mathbb{R}$ defined by (4.75) is called the* Cauchy density *on $\mathbb{R}$. The probability distribution on $\mathbb{R}$ with Cauchy density f is called the* Cauchy *probability distribution on $\mathbb{R}$.*

Example 334 (exponential density) *Assume $N = 1$, let $\lambda > 0$ and let $f_\lambda : \mathbb{R} \to \mathbb{R}$ be the function given by*

$$f_\lambda(x) \overset{def}{=} \lambda e^{-\lambda x} 1_{\mathbb{R}_+}(x), \quad \forall x \in \mathbb{R}. \tag{4.76}$$

Then, f_λ is a probability density.

Discussion. By virtue of the properties of the Lebesgue integral, we have

$$\int_{\mathbb{R}} f_\lambda(x)\, d\mu_L(x) = \int_{\mathbb{R}} \lambda e^{-\lambda x} 1_{\mathbb{R}_+}(x)\, d\mu_L(x) = \lambda \int_{\mathbb{R}_+} e^{-\lambda x} d\mu_L(x).$$

On the other hand, the integrand function is positive and continuous on $\mathbb{R}_+$. Hence, the Lebesgue integral $\int_{\mathbb{R}_+} e^{-\lambda x}\, d\mu_L(x)$ coincides with the improper Riemann integral $\int_0^{+\infty} e^{-\lambda x}\, dx$. We Then, have

$$\begin{aligned}\int_{\mathbb{R}} f_\lambda(x)\ d\mu_L(x) &= \lambda \int_0^{+\infty} e^{-\lambda x} dx = \lambda \lim_{x\to+\infty} \int_0^x e^{-\lambda u} du \\ &= \lim_{x\to+\infty} -\int_0^{-\lambda x} e^v dv = \lim_{x\to+\infty} \int_{-\lambda x}^0 e^v dv \\ &= \lim_{u\to+\infty} \left(1 - e^{-\lambda x}\right) = 1,\end{aligned}$$

as desired. □

Definition 335 (exponential density) *The probability density $f_\lambda : \mathbb{R} \to \mathbb{R}$ defined by (4.76) is called the* exponential density *on $\mathbb{R}$ with* rate *or* intensity *parameter λ. The probability distribution with exponential density f_λ is called the* exponential *probability distribution on $\mathbb{R}$.*

Remark 336 (exponential density) *The exponential density on $\mathbb{R}$ is also often expressed in terms of the parameter $\theta \equiv 1/\lambda$, which is called the* scale *or* survival *parameter. In terms of θ, the exponential density becomes the function $f_\theta : \mathbb{R} \to \mathbb{R}$ given by*

$$f_\theta(x) \overset{def}{=} \frac{1}{\theta} e^{-\frac{x}{\theta}} 1_{\mathbb{R}_+}(x), \quad \forall x \in \mathbb{R}. \tag{4.77}$$

Example 337 (Laplace density) *Assume $N = 1$, let $\mu, \sigma \in \mathbb{R}$ such that $\sigma > 0$, and let $f_{\mu,\sigma} : \mathbb{R} \to \mathbb{R}$ be the function given by*

$$f_{\mu,\sigma}(x) \overset{def}{=} \frac{1}{2\sigma} e^{-\frac{|x-\mu|}{\sigma}}, \quad \forall x \in \mathbb{R}. \tag{4.78}$$

Then, $f_{\mu,\sigma}$ is a probability density on $\mathbb{R}$.

Discussion. Since the function $f_{\mu,\sigma}$ is positive and continuous on $\mathbb{R}$, similarly to Example 332, the Lebesgue integral $\int_{\mathbb{R}} e^{-\frac{|x-\mu|}{\sigma}} d\mu(x)$ coincides with the improper Riemann integral $\int_{-\infty}^{+\infty} e^{-\frac{|x-\mu|}{\sigma}} dx$. Now, a straightforward computation gives

$$\begin{aligned}\int_{-\infty}^{+\infty} e^{-\frac{|x-\mu|}{\sigma}} dx &= \lim_{x\to+\infty} \int_{-x}^{x} e^{-\frac{|u-\mu|}{\sigma}} du = \lim_{x\to+\infty} \left(\int_{-x}^{\mu} e^{-\frac{\mu-u}{\sigma}} du + \int_{\mu}^{x} e^{-\frac{u-\mu}{\sigma}} du \right) \\ &= \lim_{x\to+\infty} \sigma \left(-\int_{\frac{\mu+x}{\sigma}}^{0} e^{-v} dv + \int_0^{\frac{\mu+x}{\sigma}} e^{-v} dv \right) = 2\sigma \lim_{x\to+\infty} \int_0^{\frac{\mu+u}{\sigma}} e^{-v} dv \\ &= 2\sigma \lim_{x\to+\infty} \left(1 - e^{-\frac{\mu+x}{\sigma}}\right) = 2\sigma.\end{aligned}$$

On account of this, it clearly follows

$$\int_{\mathbb{R}} f_{\mu,\sigma}(x)\, dx = \int_{\mathbb{R}} \frac{1}{2\sigma} e^{-\frac{|x-\mu|}{\sigma}} d\mu(x) = \frac{1}{2\sigma} \int_{\mathbb{R}} e^{-\frac{|x-\mu|}{\sigma}} d\mu(x) = \frac{1}{2\sigma} \int_{-\infty}^{+\infty} e^{-\frac{|x-\mu|}{\sigma}} dx = 1,$$

as desired. □

Definition 338 (Laplace density) *The probability density $f_{\mu,\sigma} : \mathbb{R} \to \mathbb{R}$ defined by (4.78) is called the* Laplace density *on $\mathbb{R}$ with* location parameter *μ and* scale parameter *σ on $\mathbb{R}$. The probability distribution with Laplace density $f_{\mu,\sigma}$ is called the* Laplace distribution *on $\mathbb{R}$.*

Example 339 (Gaussian density on $\mathbb{R}$) *Assume $N = 1$, let $\mu, \sigma \in \mathbb{R}$ such that $\sigma > 0$, and let $f_{\mu,\sigma} : \mathbb{R} \to \mathbb{R}$ be the function given by*

$$f_{\mu,\sigma}(x) \overset{def}{=} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad \forall x \in \mathbb{R}. \tag{4.79}$$

Then, $f_{\mu,\sigma}$ is a probability density on $\mathbb{R}$.

Discussion. Since the function $f_{\mu,\sigma}$ is positive and continuous on $\mathbb{R}$, as in Example **??**, the Lebesgue integral $\int_{\mathbb{R}} f_{\mu,\sigma}(x)\, d\mu(x)$ coincides with the improper Riemann integral $\int_{-\infty}^{+\infty} f_{\mu,\sigma}(x)\, dx$. Now, setting $y \equiv \frac{x-\mu}{\sqrt{2}\sigma}$, we have

$$\int_{-\infty}^{+\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \sqrt{2}\sigma \int_{-\infty}^{+\infty} e^{-y^2} dy.$$

Therefore, if we proved that

$$\int_{-\infty}^{+\infty} e^{-y^2} dy = \sqrt{\pi}, \tag{4.80}$$

it would follow

$$\int_{\mathbb{R}} f_{\mu,\sigma}(x)\, dx = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \frac{1}{\sqrt{2\pi}\sigma} \sqrt{2}\sigma \int_{-\infty}^{+\infty} e^{-y^2} dy = 1,$$

which is the desired result. Now, to obtain (4.80), observe first that we can write

$$\int_{-\infty}^{+\infty} e^{-y^2} dy = \lim_{r \to +\infty} \int_{-r}^{r} e^{-y^2} dy.$$

Hence,

$$\begin{aligned} \left(\int_{-\infty}^{+\infty} e^{-y^2} dy \right)^2 &= \lim_{r \to +\infty} \int_{-r}^{r} e^{-x^2} dx \int_{-r}^{r} e^{-y^2} dy \\ &= \lim_{r \to +\infty} \int_{-r}^{r} \int_{-r}^{r} e^{-x^2} e^{-y^2} dx dy = \lim_{r \to +\infty} \int_{-r}^{r} \int_{-r}^{r} e^{-\left(x^2+y^2\right)} dx dy. \end{aligned}$$

Second, writing $D(0; r)$ [resp. $D\left(0; \sqrt{2}r\right)$] for the disk centered in $0 \equiv (0, 0)$ with radius r [resp. $\sqrt{2}r$], we clearly have

$$\int_{D(0;r)} e^{-\left(x^2+y^2\right)} dx dy \leq \int_{-r}^{r} \int_{-r}^{r} e^{-\left(x^2+y^2\right)} dx dy \leq \int_{D\left(0;\sqrt{2}r\right)} e^{-\left(x^2+y^2\right)} dx dy. \tag{4.81}$$

Third, changing rectangular coordinates to polar coordinates, that is to say setting

$$x = \rho \cos(\theta), \qquad y = \rho \sin(\theta),$$

we can compute easily

$$\int_{D(0;r)} e^{-\left(x^2+y^2\right)}\,dxdy = \int_0^{2\pi} d\theta \int_0^r \rho e^{-\rho^2}\,d\rho = \pi(1-e^{-r^2}), \tag{4.82}$$

and

$$\int_{D\left(0;\sqrt{2}r\right)} e^{-\left(x^2+y^2\right)}\,dxdy = \int_0^{2\pi} d\theta \int_0^{\sqrt{2}r} \rho e^{-\rho^2}\,d\rho = \pi(1-e^{-2r^2}). \tag{4.83}$$

Finally, combining (4.81)-(4.83), and letting r go to $+\infty$, we obtain

$$\lim_{r\to+\infty} \int_{-r}^{r}\int_{-r}^{r} e^{-\left(x^2+y^2\right)}\,dxdy = \pi,$$

which clearly implies (4.80). □

Definition 340 (Gaussian density on $\mathbb{R}$) *The probability density $f_{\mu,\sigma} : \mathbb{R} \to \mathbb{R}$ defined by (4.79) is called the* Gaussian *or* normal density *on $\mathbb{R}$ with* location parameter μ *and* scale parameter σ. *The probability distribution with Gaussian density $f_{\mu,\sigma}$ is called the* Gaussian *or* normal distribution *on $\mathbb{R}$.*

Example 341 (Gaussian density on $\mathbb{R}^N$) *Assume $N > 1$, let $\mu \equiv (\mu_1, \ldots, \mu_N) \in \mathbb{R}^N$, let $\Sigma^2 \equiv \left(\sigma_{m,n}\right)_{m,n=1}^N \in \mathbb{R}^N \times \mathbb{R}^N$ such that Σ^2 is symmetric and positive definite, and let $f_{\mu,\Sigma^2} : \mathbb{R}^N \to \mathbb{R}$ be the function given by*

$$f_{\mu,\Sigma^2}(x) \overset{def}{=} \frac{1}{(2\pi)^{N/2}\det\left(\Sigma^2\right)^{1/2}} e^{-\frac{1}{2}(x-\mu)^\intercal\left(\Sigma^2\right)^{-1}(x-\mu)}, \quad \forall x \equiv (x_1, \ldots, x_N) \in \mathbb{R}^N, \tag{4.84}$$

where $\det : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ is the determinant function on $\mathbb{R}^N \times \mathbb{R}^N$ and $\left(\Sigma^2\right)^{-1}$ is the inverse of Σ^2. Then, f_{μ,Σ^2} is a probability density on $\mathbb{R}^N$.

Definition 342 (Gaussian density on $\mathbb{R}^N$) *The probability density $f_{\mu,\Sigma^2} : \mathbb{R}^N \to \mathbb{R}$ defined by (4.84) is called the* Gaussian *or* normal density *on $\mathbb{R}^N$ with* mean vector parameter μ *and* variance-covariance matrix parameter Σ. *The probability of Gaussian density f_{μ,Σ^2} is called the* Gaussian *or* normal distribution *on $\mathbb{R}^N$.*

Example 343 (lognormal density on $\mathbb{R}$) *Assume $N = 1$, let $\mu, \sigma \in \mathbb{R}$ such that $\sigma > 0$, and let $f_{\mu,\sigma} : \mathbb{R} \to \mathbb{R}$ be the function given by*

$$f_{\mu,\sigma}(x) \overset{def}{=} \frac{1}{\sqrt{2\pi}\sigma x} \exp\left(-\frac{\left(\ln(x)-\mu\right)^2}{2\sigma^2}\right) 1_{\mathbb{R}_+}(x), \quad \forall x \in \mathbb{R}. \tag{4.85}$$

Then, $f_{\mu,\sigma}$ is a probability density on $\mathbb{R}$.

Discussion. We have

$$\begin{aligned}\int_{\mathbb{R}} f_{\mu,\sigma}(x)\,d\mu(x) &= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}\sigma x} \exp\left(-\frac{\left(\ln(x)-\mu\right)^2}{2\sigma^2}\right) 1_{\mathbb{R}_+}(x)\,d\mu(x) \\ &= \frac{1}{\sqrt{2\pi}\sigma}\int_{\mathbb{R}_+} \frac{1}{x} \exp\left(-\frac{\left(\ln(x)-\mu\right)^2}{2\sigma^2}\right) d\mu(x).\end{aligned} \tag{4.86}$$

Furthermore, since the integrand function is positive and continuous on $\mathbb{R}_+$, we have

$$\int_{\mathbb{R}_+} \frac{1}{x} \exp\left(-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right) d\mu(x) = \int_0^{+\infty} \frac{1}{x} \exp\left(-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right) dx. \tag{4.87}$$

Now, setting $y \equiv \frac{\ln(x)-\mu}{\sqrt{2}\sigma}$, we have $x = e^{\left(\sqrt{2}\sigma y+\mu\right)}$ and $dx = e^{\left(\sqrt{2}\sigma y+\mu\right)}\sqrt{2}\sigma dy$. In addition, as $x \to 0^+$ [resp. $x \to +\infty$] we have $y \to -\infty$ [resp. $y \to +\infty$]. Considering (4.80), it Then, follows,

$$\int_0^{+\infty} \frac{1}{x} \exp\left(-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right) dx = \sqrt{2}\sigma \int_{-\infty}^{+\infty} e^{-y^2} dy = \sqrt{2\pi}\sigma. \tag{4.88}$$

Combining (4.86)-(4.88), we obtain

$$\int_{\mathbb{R}} f_{\mu,\sigma}(x)\, d\mu(x) = 1,$$

which shows the claim. □

Definition 344 (lognormal density on $\mathbb{R}$) *The probability density $f_{\mu,\sigma} : \mathbb{R} \to \mathbb{R}$ defined by (4.85) is called the* lognormal density *on $\mathbb{R}$ with* location parameter μ *and* scale parameter σ. *The probability with lognormal density $f_{\mu,\sigma}$ is called the* lognormal distribution *on $\mathbb{R}$.*

Lemma 345 (Gamma function) *We have*

$$\int_{\mathbb{R}_+} x^{\alpha-1} e^{-x} d\mu(x) < +\infty \tag{4.89}$$

for every $\alpha > 0$.

Proof. Since the integrand function is positive and continuous on $\mathbb{R}_+$, we can write

$$\int_{\mathbb{R}_+} x^{\alpha-1} e^{-x} d\mu(x) = \int_0^{+\infty} x^{\alpha-1} e^{-x} dx = \lim_{x\to 0^+} \int_x^1 u^{\alpha-1} e^{-u} du + \lim_{x\to+\infty} \int_1^x u^{\alpha-1} e^{-u} du$$

Now,

$$u^{\alpha-1} e^{-u} \le u^{\alpha-1}, \quad \forall x \in (0,1].$$

Therefore,

$$\lim_{x\to 0^+} \int_x^1 u^{\alpha-1} e^{-u}\, du \le \lim_{x\to 0^+} \int_x^1 u^{\alpha-1}\, du = \lim_{x\to 0^+} \frac{1}{\alpha} u^{\alpha} \Big|_x^1 = \frac{1}{\alpha}.$$

In addition,

$$u^{\alpha-1} e^{-u} \le u^{\lfloor\alpha\rfloor} e^{-u}, \quad \forall u \in [1,+\infty),$$

where $\lfloor\cdot\rfloor : \mathbb{R} \to \mathbb{R}$ is the floor function. Hence,

$$\lim_{x\to+\infty} \int_1^x u^{\alpha-1} e^{-u}\, du \le \lim_{x\to+\infty} \int_1^x u^{\lfloor\alpha\rfloor} e^{-u} du.$$

On the other hand, repeatedly integrating by parts,

$$\begin{aligned}\int_1^x u^{\lfloor\alpha\rfloor}e^{-u}du &= -\int_1^x u^{\lfloor\alpha\rfloor}de^{-u} = -\,u^{\lfloor\alpha\rfloor}e^{-u}\big|_1^x + \lfloor\alpha\rfloor\int_1^x u^{\lfloor\alpha\rfloor-1}e^{-u}du\\ &= -\,u^{\lfloor\alpha\rfloor}e^{-u}\big|_1^x - \lfloor\alpha\rfloor\,u^{\lfloor\alpha\rfloor-1}e^{-u}\big|_1^x + \lfloor\alpha\rfloor\left(\lfloor\alpha\rfloor-1\right)\int_1^x u^{\lfloor\alpha\rfloor-2}e^{-u}\,du\\ &= \dots\\ &= -\,u^{\lfloor\alpha\rfloor}e^{-u}\big|_1^x - \lfloor\alpha\rfloor\,u^{\lfloor\alpha\rfloor-1}e^{-u}\big|_1^x - \lfloor\alpha\rfloor\left(\lfloor\alpha\rfloor-1\right)u^{\lfloor\alpha\rfloor-2}e^{-u}\big|_1^x - \cdots - \lfloor\alpha\rfloor!\,e^{-u}\big|_1^x\\ &= \left(1+\lfloor\alpha\rfloor+\lfloor\alpha\rfloor\left(\lfloor\alpha\rfloor-1\right)+\cdots+\lfloor\alpha\rfloor!\right)e^{-1}\\ &\quad-\left(x^{\lfloor\alpha\rfloor}+\lfloor\alpha\rfloor\,x^{\lfloor\alpha\rfloor-1}+\lfloor\alpha\rfloor\left(\lfloor\alpha\rfloor-1\right)x^{\lfloor\alpha\rfloor-1}+\cdots+\lfloor\alpha\rfloor!\right)e^{-x}.\end{aligned}$$

Therefore,

$$\begin{aligned}\lim_{x\to+\infty}\int_1^x u^{\alpha-1}e^{-u}du &\leq \lim_{x\to+\infty}\int_1^x u^{\lfloor\alpha\rfloor}e^{-u}\,du\\ &= \left(1+\lfloor\alpha\rfloor+\lfloor\alpha\rfloor\left(\lfloor\alpha\rfloor-1\right)+\cdots+\lfloor\alpha\rfloor!\right)e^{-1}\\ &\quad-\lim_{x\to+\infty}\left(x^{\lfloor\alpha\rfloor}+\lfloor\alpha\rfloor\,x^{\lfloor\alpha\rfloor-1}+\lfloor\alpha\rfloor\left(\lfloor\alpha\rfloor-1\right)x^{\lfloor\alpha\rfloor-1}+\cdots+\lfloor\alpha\rfloor!\right)e^{-x}\\ &= \left(1+\lfloor\alpha\rfloor+\lfloor\alpha\rfloor\left(\lfloor\alpha\rfloor-1\right)+\cdots+\lfloor\alpha\rfloor!\right)e^{-1}.\end{aligned}$$

From what shown above, it follows

$$\int_0^{+\infty} x^{\alpha-1}e^{-x}dx \leq \frac{1}{\alpha}+\left(1+\lfloor\alpha\rfloor+\lfloor\alpha\rfloor\left(\lfloor\alpha\rfloor-1\right)+\cdots+\lfloor\alpha\rfloor!\right)e^{-1},$$

for every $\alpha>0$, which is the desired result. □

Definition 346 (Gamma function) *We call the function $\Gamma:\mathbb{R}_{++}\to\mathbb{R}_{++}$ given by*

$$\Gamma(\alpha)\overset{def}{=}\int_{\mathbb{R}_+} x^{\alpha-1}e^{-x}d\mu(x),\quad\forall\alpha>0, \tag{4.90}$$

the Gamma function*.*

Proposition 347 (Gamma funcion) *We have*

$$\Gamma(\alpha+1)=\alpha\Gamma(\alpha),\quad\forall\alpha>0. \tag{4.91}$$

In particular,

$$\Gamma(n+1)=n!,\quad\forall n\in\mathbb{N}. \tag{4.92}$$

Proof. By Equation (4.90), we can write

$$\begin{aligned}\Gamma(\alpha+1) &= \int_{\mathbb{R}_+} x^{\alpha}e^{-x}d\mu(x) = \int_0^{+\infty} x^{\alpha}e^{-x}dx = \lim_{x\to+\infty}\int_0^x u^{\alpha}e^{-u}du\\ &= \lim_{x\to+\infty}\left(-\int_0^x u^{\alpha}de^{-u}\right) = \lim_{x\to+\infty}\left(-\,u^{\alpha}e^{-u}\big|_0^x + \int_0^x \alpha u^{\alpha-1}e^{-u}du\right)\\ &= \alpha\lim_{x\to+\infty}\int_0^x u^{\alpha-1}e^{-u}du = \alpha\int_0^{+\infty} x^{\alpha-1}e^{-x}dx = \alpha\int_{\mathbb{R}_+} x^{\alpha-1}e^{-x}d\mu(x)\\ &= \alpha\Gamma(\alpha),\end{aligned}$$

which proves the desired (4.91). In particular, since

$$\begin{aligned}\Gamma(1) &= \int_{\mathbb{R}_+} e^{-x} d\mu(x) = \int_0^{+\infty} e^{-x} dx = \lim_{x\to+\infty} \int_0^x e^{-u} du \\ &= \lim_{x\to+\infty} \left(-\int_0^{-x} e^u du \right) = \lim_{x\to+\infty} \int_{-x}^0 e^u du = \lim_{x\to+\infty} e^u|_{-x}^0 \\ &= 1 = 0!,\end{aligned}$$

thanks to (4.91) and the Induction Principle, we obtain

$$\Gamma(n+1) = n\Gamma(n) = n(n-1)! = n!.$$

This yields (4.92). □

Corollary 348 (Gamma function) *We have*

$$\Gamma(\alpha+1) = \alpha(\alpha-1)\cdots(\alpha-\lfloor\alpha\rfloor)\Gamma(\alpha-\lfloor\alpha\rfloor), \tag{4.93}$$

for every $\alpha > 0$, where $[\alpha]$ is the integer part of α.

Proposition 349 (Gamma function) *We have*

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}. \tag{4.94}$$

Proof. In fact,

$$\Gamma\left(\frac{1}{2}\right) = \int_{\mathbb{R}_+} x^{-\frac{1}{2}} e^{-x} d\mu(x) = \int_0^{+\infty} x^{-\frac{1}{2}} e^{-x} dx.$$

On the other hand, setting $x \equiv \frac{1}{2}y^2$, we obtain

$$\begin{aligned}\int_0^{+\infty} x^{-\frac{1}{2}} e^{-x} dx &= \int_0^{+\infty} \left(\frac{1}{2}y^2\right)^{-\frac{1}{2}} e^{-\frac{1}{2}y^2} d\left(\frac{1}{2}y^2\right) = \sqrt{2}\int_0^{+\infty} e^{-\frac{1}{2}y^2} dy \\ &= 2\sqrt{\pi}\left(\frac{1}{\sqrt{2\pi}}\int_0^{+\infty} e^{-\frac{1}{2}y^2} dy\right) = \sqrt{\pi}\left(\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-\frac{1}{2}y^2} dy\right) \\ &= \sqrt{\pi},\end{aligned}$$

as desired. □

Example 350 (Gamma density) *Let $\alpha, \theta > 0$ and let $f_{\alpha,\theta} : \mathbb{R} \to \mathbb{R}$ be the function given by*

$$f_{\alpha,\theta}(x) \stackrel{def}{=} \frac{1}{\theta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\theta}} 1_{\mathbb{R}_+}(x), \quad \forall x \in \mathbb{R}. \tag{4.95}$$

Then, $f_{\alpha,\theta}$ is a probability density on $\mathbb{R}$.

Proof. As in Example 332, the Lebesgue integral $\int_{\mathbb{R}} f_{\alpha,\theta}(x)\, d\mu_L(x)$ becomes the improper Riemann integral $\int_{-\infty}^{+\infty} f_{\alpha,\theta}(x)\, dx$. Now, thanks to Equation (4.90), we have

$$\int_{-\infty}^{+\infty} f_{\alpha,\theta}(x)\, dx = \frac{1}{\Gamma(\alpha)} \int_{-\infty}^{+\infty} \left(\frac{x}{\theta}\right)^{\alpha-1} e^{-\frac{x}{\theta}}\, d\left(\frac{x}{\theta}\right) = \frac{1}{\Gamma(\alpha)} \int_{-\infty}^{+\infty} y^{\alpha-1} e^{-y}\, dy = 1,$$

which implies the desired result. □

Definition 351 (Gamma density) *The probability density $f_{\alpha,\theta} : \mathbb{R} \to \mathbb{R}$ given by (4.95) is referred to as the* Gamma density *with* shape *parameter α and* scale *parameter θ. In particular, when $\theta = 1$ we speak of* standard Gamma density*. The probability distribution with (standard) Gamma density is referred to as the (*standard*)* Gamma distribution*.*

Remark 352 (Gamma density) *The parameter θ is called the* scale *parameter, because the smaller [resp. larger] is θ the more compressed [resp. stretched] is the Gamma density along the x-axis.*

Remark 353 (Gamma density) *The Gamma density is often expressed in terms of the parameter $\lambda \equiv 1/\theta$, which is called the* rate *or* inverse scale *parameter. Clearly, the smaller [resp. larger] is λ the more stretched [resp. compressed] is the Gamma density along the x-axis. In terms of the parameter λ, the Gamma density becomes the function $f_{\alpha,\lambda} : \mathbb{R} \to \mathbb{R}$ given by*

$$f_{\alpha,\lambda}(x) \stackrel{def}{=} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} 1_{\mathbb{R}_+}(x), \quad \forall x \in \mathbb{R}. \tag{4.96}$$

Remark 354 (Gamma density) *The Gamma density with shape parameter $\alpha = 1$, reduces to the exponential density.*

Example 355 (chi-square density) *Let $\nu > 0$ and let $f_\nu : \mathbb{R} \to \mathbb{R}$ be the function given by*

$$f_\nu(x) \stackrel{def}{=} \frac{1}{2^{\frac{\nu}{2}} \Gamma\left(\frac{\nu}{2}\right)} x^{\frac{\nu}{2}-1} e^{-\frac{x}{2}} 1_{\mathbb{R}_+}(x), \quad \forall x \in \mathbb{R}. \tag{4.97}$$

Then, f_ν is a probability density.

Definition 356 (chi-square density) *The probability density $f_\nu : \mathbb{R} \to \mathbb{R}$ given by 4.97 is referred to as the* chi-square density *with ν degrees of freedom[2]. The probability distribution with chi-square density is called the* chi-square distribution*.*

Remark 357 (chi-square density) *The chi-square density with ν degrees of freedom concides with the Gamma density $f_{\alpha,\theta} : \mathbb{R} \to \mathbb{R}$ of shape parameter $\alpha \equiv \nu/2$ and scale parameter $\theta \equiv 2$.*

Remark 358 (chi-square density) *The chi-square density with $\nu = 2$ degrees of freedom coincides with the exponential density $f_\lambda : \mathbb{R} \to \mathbb{R}$ of rate parameter $\lambda = 1/2$.*

[2]Despite the chi–square density is defined for any $\nu > 0$, only integer values of ν are of simple statistical interpretation.

Example 359 (Student's t density) *Let $\nu > 0$ and let $f_\nu : \mathbb{R} \to \mathbb{R}$ be the function given by*

$$f_\nu(x) \overset{def}{=} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad \forall x \in \mathbb{R}_+ \tag{4.98}$$

Then, f_ν is a probability density.

Definition 360 (Student's t density) *The probability density $f_\nu : \mathbb{R} \to \mathbb{R}$ given by 4.98 is referred to as the* Student's t-density *with ν degrees of freedom*[3]*. The probability distribution with Student's t density is called the* Student's t-distribution*.*

Example 361 (F-density) *Let $\nu, \delta > 0$ and let $f_{\nu,\delta} : \mathbb{R} \to \mathbb{R}$ be the function given by*

$$f_{\nu,\delta}(x) \overset{def}{=} \frac{\nu}{\delta} \frac{\Gamma\left(\frac{\nu+\delta}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\Gamma\left(\frac{\delta}{2}\right)} \frac{\left(\frac{\nu}{\delta}x\right)^{\frac{\nu}{2}-1}}{\left(1+\frac{\nu}{\delta}x\right)^{\frac{\nu+\delta}{2}}} 1_{\mathbb{R}_+}(x), \quad \forall x \in \mathbb{R}. \tag{4.99}$$

Then, $f_{\nu,\delta}$ is a probability density.

Definition 362 (F-density) *The probability density $f_{\nu,\delta} : \mathbb{R} \to \mathbb{R}$ given by* **??** *is referred to as the* F-density *or the* Fisher-Snedecor density *with degrees of freedom ν and δ.*[4]*. The probability of F-density is referred to as the* F-distribution *or the* Fisher-Snedecor distribution*.*

The F-distribution was first derived by George Snedecor and takes its name in honor of Ronald Fisher.

[3]Despite the Student's t-density is defined for any $\nu > 0$, only integer values of ν are of simple statistical interpretation.

[4]Likewise the Student's t-density, for F-density only integer values of ν and δ are of simple statistical interpretation.

4.4 Independent Events and σ-algebras of Events

Loosely speaking, by saying that two events E and F are *independent* we mean that the probability of observing the occurrence of one of the two events does not affect the probability of observing the occurrence of the other. For instance, the probability of a newborn baby being a girl is clearly independent of the probability that the weekday of birth is Monday. On the other hand, the probability of a newborn baby being a girl is not independent of the probability that the length or weight of the infant takes value in some interval (see WHO data at`https://www.cdc.gov/growthcharts/data/who/grchrt_boys_24lw_100611.pdf` and `https://www.cdc.gov/growthcharts/data/who/GrChrt_Girls_24LW_9210.pdf`). Within the probability space model we can formalize this idea.

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space and let E and F be two events of $\mathcal{E}$.

Definition 363 *We say E and F are* independent *(with respect to $\mathbf{P}$), if*

$$\mathbf{P}(E \cap F) = \mathbf{P}(E)\,\mathbf{P}(F). \tag{4.100}$$

Example 364 *Two identical fair coins, labeled by A and B, are flipped. The events "the outcome of the flip of the coin A is heads" and "the outcome of the flip of the coin B is heads" are independent with respect to the naive probability.*

Discussion. Representing the outcomes heads an tails of the coin flips with the numbers 1 and 0, respectively, let $\Omega_A \equiv \{0,1\}$ and $\Omega_B \equiv \{0,1\}$ be the sample spaces for the flip of the coin A and B, respectively. The standard sample space for the flip of the two coins is Then, $\Omega \equiv \Omega_A \times \Omega_B = \{(0,0),(0,1),(1,0),(1,1)\}$. Since we need to distinguish the outcomes of the two flips, we have to consider the complete information, that is the algebra of events $\mathcal{E} \equiv \mathcal{P}(\Omega)$. Hence, the events "the outcome of the flip of the coin A is heads" and "the outcome of the flip of the coin B is heads" are represented by the subsets of the sample space

$$H_A \equiv \{(1,0),(1,1)\} \qquad \text{and} \qquad H_B \equiv \{(0,1),(1,1)\},$$

respectively, and the event "the outcome of the flip of the coins A and B is head" is represented by the subset

$$H_A \cap H_B = \{(1,1)\}.$$

Since the coins are assumed to be fair, there is no reason to think that different elementary events of the sample space Ω are not equally probable. This justifies the application of the naive probability on $\mathcal{E}$. It Then, follows

$$\mathbf{P}(H_j) = \frac{|H_j|}{|\Omega|} = \frac{1}{2},$$

for $j = A, B$, and

$$\mathbf{P}(H_A \cap H_B) = \frac{|H_A \cap H_B|}{|\Omega|} = \frac{1}{4}.$$

As a consequence, Equation (4.100) holds true. Note that if we consider the sequence $(q_{j,k})_{j,k=0}^{1}$ given by

$$q_{0,0} \equiv \frac{1}{6}, \quad q_{0,1} \equiv \frac{1}{4} \quad , q_{1,0} \equiv \frac{1}{4}, \quad q_{1,1} \equiv \frac{1}{3},$$

we have

$$\sum_{j,k=0}^{1} q_{j,k} = \frac{1}{6} + \frac{1}{4} + \frac{1}{4} + \frac{1}{3} = 1,$$

that is $(q_{j,k})_{j,k=0}^{1}$ is a probability distribution. Now, considering the probability $\mathbf{Q} : \mathcal{E} \to \mathbb{R}_+$ with distribution $(q_{j,k})_{j,k=0}^{1}$ (see Definition 286), we have

$$\mathbf{Q}(H_A) = \mathbf{Q}(1,0) + \mathbf{Q}(1,1) = \frac{7}{12}, \quad \mathbf{Q}(H_B) = \mathbf{Q}(1,0) + \mathbf{Q}(1,1) = \frac{7}{12},$$

and

$$\mathbf{Q}(H_A \cap H_B) = \mathbf{Q}(1,1) = \frac{1}{3}.$$

Thus, the events H_A and H_B are not independent with respect to $\mathbf{Q}$. □

Example 365 *A fair coin is flipped. If the outcome of the first flip is heads the fair coin is flipped again unaltered, but if the outcome of the first flip is tails, the coin is rigged to show heads twice often than showing tails. For this random phenomenon, the events "the outcome of the first flip is heads" and "the outcome of the second flip is heads" cannot be considered as independent.*

Discussion. As in the former example, let us consider the standard sample spaces $\Omega_1 \equiv \{0,1\}$ [resp. $\Omega_2 \equiv \{0,1\}$] to represent the first [resp. second] flip of the coin. The standard sample space for the two flips is again $\Omega \equiv \Omega_1 \times \Omega_2 = \{(0,0),(0,1),(1,0),(1,1)\}$, we have to consider again the complete information, $\mathcal{E} \equiv \mathcal{P}(\Omega)$, and the events

$$H_1 \equiv \{(1,0),(1,1)\} \qquad \text{and} \qquad H_2 \equiv \{(0,1),(1,1)\}.$$

However, in this example the coin may be rigged depending on the results of the first flip. Seeking a probability $\mathbf{Q} : \mathcal{E} \to \mathbb{R}_+$ which accounts for this random phenomenon, it seems reasonable to set the following conditions

$$\mathbf{Q}(1,1) = \frac{1}{4}, \quad \mathbf{Q}(1,0) = \frac{1}{4}, \quad \mathbf{Q}(0,1) = 2\mathbf{Q}(0,0).$$

On the other hand, we need

$$\mathbf{Q}(0,0) + \mathbf{Q}(0,1) + \mathbf{Q}(1,0) + \mathbf{Q}(1,1) = 1.$$

It Then, follows

$$\mathbf{Q}(0,0) + 2\mathbf{Q}(0,0) + \frac{1}{4} + \frac{1}{4} = 1,$$

which yields

$$\mathbf{Q}(0,0) = \frac{1}{6} \quad \text{and} \quad \mathbf{Q}(0,1) = \frac{1}{3}.$$

Setting $q_{j,k} \equiv \mathbf{Q}(j,k)$, for $j,k = 0,1$, the sequence $(q_{j,k})_{j,k=0}^{1}$ is the probability density on the sample space Ω which accounts for the possibility of rigging the coin. This is just the probability density shown in Example 364, with respect to which the events H_1 and H_2 are not independent. □

From Examples 364 and 365, it is clearly seen the crucial role of the assigned probability in establishing the independence of events: two events may be independent with respect to a probability and not independent with respect to another probability. A careful assignment of the probability that we use to build the matematical model for a random phenomenon or experiment is of the utmost importance.

Proposition 366 *The events E and F are independent if and only if the events E and F^c are. The same result holds true for any of the two couples of events E^c, F and E^c, F^c.*

Proof. Assume E and F are independent and consider the events E and F^c. Since we can write

$$\mathbf{P}(E) = \mathbf{P}(E \cap \Omega) = \mathbf{P}(E \cap (F \cup F^c)) = \mathbf{P}((E \cap F) \cup (E \cap F^c)) = \mathbf{P}(E \cap F) + \mathbf{P}(E \cap F^c), \tag{4.101}$$

we obtain

$$\mathbf{P}(E \cap F^c) = \mathbf{P}(E) - \mathbf{P}(E \cap F) = \mathbf{P}(E) - \mathbf{P}(E)\mathbf{P}(F) = \mathbf{P}(E)(1 - \mathbf{P}(F)) = \mathbf{P}(E)\mathbf{P}(F^c).$$

This yields the independence of E and F^c. Conversely, Assume E and F^c are independent and consider the events E and F. Still from (366) we obtain

$$\begin{aligned}\mathbf{P}(E \cap F) &= \mathbf{P}(E) - \mathbf{P}(E \cap F^c) = \mathbf{P}(E) - \mathbf{P}(E)\mathbf{P}(F^c) \\ &= \mathbf{P}(E) - \mathbf{P}(E)(1 - \mathbf{P}(F)) = \mathbf{P}(E)\mathbf{P}(F),\end{aligned}$$

which yields the independence of E and F. The proof of the Proposition can be completed with similar arguments. □

Example 367 *Two identical fair coins, denoted by A and B, are flipped. Show that the events "the outcome of the flip of coin A is heads" and "the outcome of the flip of coin B is tails" are independent with respect to the naive probability.*

Discussion. With reference to Example 364, the events "the outcome of the flip of coin A is heads" and "the outcome of coin B is tails" are represented by the subsets

$$H_A \equiv \{(1,0),(1,1)\} \qquad \text{and} \qquad T_B \equiv \{(0,0),(1,0)\}.$$

Recall also that the event "the outcome of the flip of coin B is heads" is represented by

$$H_B \equiv \{(0,1),(1,1)\}.$$

Now, we have

$$T_B = H_B^c$$

and, as shown in Example 364, the events H_A and H_B are independent. Thus, by virtue of Proposition 366, the events H_A and T_B are independent. □

Remark 368 *Assume $\mathbf{P}(E) = 0$ or $\mathbf{P}(F) = 0$. Then, the events E and F are independent.*

Proof. Since

$$\mathbf{P}(E \cap F) \leq \mathbf{P}(E) \quad \text{and} \quad \mathbf{P}(E \cap F) \leq \mathbf{P}(F),$$

under the considered assumption we have

$$\mathbf{P}(E \cap F) = 0.$$

It Then, clearly follows

$$\mathbf{P}(E \cap F) = \mathbf{P}(E)\,\mathbf{P}(F),$$

as desired. □

Let $(E_j)_{j\in J}$ be a subfamily of $\mathcal{E}$.

Definition 369 *We say that the events of the family $(E_j)_{j\in J}$ are* pariwise independent *(with respect to $\mathbf{P}$), if we have*

$$\mathbf{P}(E_{j_1} \cap E_{j_2}) = \mathbf{P}(E_{j_1})\,\mathbf{P}(E_{j_2}), \tag{4.102}$$

for all $j_1, j_2 \in J$, such that $j_1 \neq j_2$.

Definition 370 *We say that the events of the family $(E_j)_{j\in J}$ are (*totally *or* mutually*)* independent *(with respect to $\mathbf{P}$), if we have*

$$\mathbf{P}(E_{j_1} \cap \cdots \cap E_{j_n}) = \mathbf{P}(E_{j_1}) \cdots \mathbf{P}(E_{j_n}), \tag{4.103}$$

for every finite subset $\{j_1, \ldots, j_n\}$ of J.

Remark 371 *If the events of the family $(E_j)_{j\in J}$ are totally independent, Then, they are also pairwise independent. The converse is not true.*

Example 372 *With reference to the probability space $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$, where $\Omega \equiv \{1,2,3,4\}$, $\mathcal{E} \equiv \mathcal{P}(\Omega)$, and $\mathbf{P} : \Omega \to \mathbb{R}_+$ is the naive probability consider the events $E_1 \equiv \{1,2\}$, $E_2 \equiv \{2,3\}$, $E_3 \equiv \{1,3\}$. Then, the events E_1, E_2, and E_3 are pairwise independent, but not totally independent.*

Discussion. We have

$$\mathbf{P}(E_j) = \frac{|E_j|}{|\Omega|} = \frac{1}{2},$$

for every $j = 1,2,3$. Furthermore,

$$\mathbf{P}(E_j \cap E_k) = \frac{|E_j \cap E_k|}{|\Omega|} = \frac{1}{4}$$

for all $j, k = 1,2,3$ such that $j \neq k$. Therefore,

$$\mathbf{P}(E_j \cap E_k) = \mathbf{P}(E_j)\,\mathbf{P}(E_k)$$

for all $j, k = 1,2,3$ such that $j \neq k$. This shows that the events of the family $(E_j)_{j=1}^3$ are pairwise independent. On the other hand, we have

$$E_1 \cap E_2 \cap E_3 = \varnothing.$$

Hence,

$$\mathbf{P}(E_1 \cap E_2 \cap E_3) = 0 \neq \frac{1}{8} = \mathbf{P}(E_1)\,\mathbf{P}(E_2)\,\mathbf{P}(E_3).$$

This implies that the events of the family $(E_j)_{j=1}^3$ are not totally independent. □

Example 373 *With reference to the probability space* $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$*, where* $\Omega \equiv \{1, 2, \ldots, 7, 8\}$*,* $\mathcal{E} \equiv \mathcal{P}(\Omega)$*, and* $\mathbf{P} : \mathcal{E} \to \mathbb{R}_+$ *is the naive probability, consider the events* $E_1 \equiv \{1, 2, 3, 4\}$*,* $E_2 \equiv \{1, 2, 5, 6\}$*,* $E_3 \equiv \{1, 3, 7, 8\}$*. Then, despite*

$$\mathbf{P}(E_1 \cap E_2 \cap E_3) = \mathbf{P}(E_1)\,\mathbf{P}(E_2)\,\mathbf{P}(E_3),$$

the events of the family $(E_j)_{j=1}^3$ *are not totally independent because two of them are not pairwise independent.*

Discussion. Similarly to Example 372, we have

$$\begin{array}{l} \mathbf{P}(E_1) = \mathbf{P}(E_2) = \mathbf{P}(E_3) = \frac{1}{2}, \\ \mathbf{P}(E_1 \cap E_2 \cap E_3) = \frac{1}{8}, \\ \mathbf{P}(E_1 \cap E_2) = \mathbf{P}(E_1 \cap E_3) = \frac{1}{4}, \\ \mathbf{P}(E_2 \cap E_3) = \frac{1}{8}. \end{array} \tag{4.104}$$

From (4.104) it Then, follows

$$\mathbf{P}(E_1 \cap E_2 \cap E_3) = \mathbf{P}(E_1)\,\mathbf{P}(E_2)\,\mathbf{P}(E_3)$$

and

$$\mathbf{P}(E_1 \cap E_2) = \mathbf{P}(E_1)\,\mathbf{P}(E_2), \quad \mathbf{P}(E_1 \cap E_3) = \mathbf{P}(E_1)\,\mathbf{P}(E_3).$$

On the other hand,

$$\mathbf{P}(E_2 \cap E_3) = \frac{1}{8} \neq \frac{1}{4} = \mathbf{P}(E_2)\,\mathbf{P}(E_3).$$

This prevents the total independence of the events in $(E_j)_{j=1}^3$. □

Example 374 *With reference to the probability space* $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$*, where* $\Omega \equiv \{1, 2, \ldots, 26, 27\}$*,* $\mathcal{E} \equiv \mathcal{P}(\Omega)$*, and* $\mathbf{P} : \mathcal{E} \to \mathbb{R}_+$ *is the naive probability, consider the events* $E_1 \equiv \{1, 2, \ldots, 8, 9\}$*,* $E_2 \equiv \{1, 10, \ldots, 16, 17\}$*,* $E_3 \equiv \{1, 18, \ldots, 24, 25\}$*. Then, despite*

$$\mathbf{P}(E_1 \cap E_2 \cap E_3) = \mathbf{P}(E_1)\,\mathbf{P}(E_2)\,\mathbf{P}(E_3),$$

the events of the family $(E_j)_{j=1}^3$ *are not totally independent because none of them are pairwise independent.*

Discussion. Similarly to Examples 372 and 374, we have

$$\begin{array}{l} \mathbf{P}(E_1) = \mathbf{P}(E_2) = \mathbf{P}(E_3) = \frac{1}{3}, \\ \mathbf{P}(E_1 \cap E_2 \cap E_3) = \frac{1}{27}, \\ \mathbf{P}(E_1 \cap E_2) = \mathbf{P}(E_1 \cap E_3) = \mathbf{P}(E_1 \cap E_3) = \frac{1}{27}. \end{array} \tag{4.105}$$

From (4.105) it Then, follows

$$\mathbf{P}(E_1 \cap E_2 \cap E_3) = \mathbf{P}(E_1)\,\mathbf{P}(E_2)\,\mathbf{P}(E_3).$$

and

$$\mathbf{P}(E_j \cap E_k) \neq \mathbf{P}(E_j)\,\mathbf{P}(E_k),$$

for all $j, k = 1, 3$, such that $j \neq k$. □

Example 375 *With reference to Example 301, we use the notion of independent event to derive Equation (4.51).*

Discussion. Denote by W_h [resp. B_j] the event "the outcome of the hth [resp. jth] draw is a white [resp. black] ball", where the index h [resp. j] takes k [resp. $n-k$] different values in $\{1,\ldots,n\}$. According to this model, the event "the outcome of the h_1th,...,h_kth draw is a white ball and the outcome of the j_1th,...,j_{n-k}th draw is a black ball" is represented by

$$W_{h_1} \cap \ldots \cap W_{h_k} \cap B_{j_1} \cap \ldots \cap B_{j_{n-k}},$$

where $\{h_1,...,h_k\} \subseteq \{1,...,n\}$ and $\{j_1,...,j_{n-k}\} = \{1,...,n\} - \{h_1,...,h_k\}$. Since after any draw the drawn ball is replaced into the urn, the composition of the urn does not change draw after draw. Then, applying the naive probability, we obtain

$$\mathbf{P}(W_h) = p, \quad \mathbf{P}(B_j) = q,$$

for all $h, j = 1,\ldots,n$. In addition, the structure of the random phenomenon leads to think that the events $W_{h_1},\ldots,W_{h_k},B_{j_1},\ldots,B_{j_{n-k}}$ can be assumed totally independent. Therefore,

$$\mathbf{P}\left(W_{h_1} \cap \ldots \cap W_{h_k} \cap B_{j_1} \cap \ldots \cap B_{j_{n-k}}\right) = \mathbf{P}(W_{h_1}) \cdots \mathbf{P}(W_{h_k}) \cdots \mathbf{P}(B_{j_1}) \cdots \mathbf{P}\left(B_{j_{n-k}}\right) = p^k q^{n-k}.$$

In the end, since in our problem the position according to which the white balls are sampled in n draws is irrelevant, and only the number of the drawn white ball matters, the desired result clearly follows by summing on all possible subsets of k elements from the set $\{1,...,n\}$ (see Example 301). □

Example 376 *Consider an urn containing $N \geq 3$ balls of $m \geq 3$ different colors of which M_1 are white, M_2 are red, ..., and M_m are black, where $1 \leq M_\ell$, for $\ell = 1,\ldots,m$, and $\sum_{\ell=1}^m M_\ell = N$. Write*

$$p_\ell \equiv \frac{M_\ell}{N}, \quad \forall \ell = 1,\ldots,m.$$

Draw n balls from the urn with replacement *of the drawn ball in the urn. The probability $p_{n,k_1,\ldots,k_m}$ that the drawn sample of n balls contains k_1 white balls, k_2 red balls, ..., k_m black balls, where $k_\ell \geq 0$, for every $\ell = 1,\ldots,m$, and $\sum_{\ell=1}^m k_\ell = n$, is given by the multinomial density (4.53).*

Discussion. . □

The notion of independence can be easily extended to families of events.
Let $\mathcal{F}$ and $\mathcal{G}$ two sub-σ-algebras of $\mathcal{E}$.

Definition 377 *We say that $\mathcal{F}$ and $\mathcal{G}$ are* independent *(with respect to $\mathbf{P}$), if every event F of $\mathcal{F}$ is independent of every event G of $\mathcal{G}$. In symbols,*

$$\mathbf{P}(F \cap G) = \mathbf{P}(F)\,\mathbf{P}(G), \tag{4.106}$$

for every $F \in \mathcal{F}$ and $G \in \mathcal{G}$.

Let $(\mathcal{F}_j)_{j\in J}$ be a collection of sub-σ-algebras of $\mathcal{E}$.

Definition 378 *We say that the σ-algebras of the collection $(\mathcal{F}_j)_{j\in J}$ are* pairwise independent *(with respect to $\mathbf{P}$), if for all $j_1, j_2 \in J$, such that $j_1 \neq j_2$, the σ-algebras $\mathcal{F}_{j_1}$ and $\mathcal{F}_{j_2}$ are independent.*

Definition 379 *We say that the σ-algebras of the collection $(\mathcal{F}_j)_{j\in J}$ are (*totally*)* independent *(with respect to $\mathbf{P}$), if for every finite subset $\{j_1, \ldots, j_n\}$ of J we have*

$$\mathbf{P}(F_{j_1} \cap \cdots \cap F_{j_n}) = \mathbf{P}(F_{j_1}) \cdot \cdots \cdot \mathbf{P}(F_{j_n}),$$

for every $F_{j_k} \in \mathcal{F}_{j_k}$, on varing of $k = 1, \ldots, n$.

4.5 Conditional Probabilities

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space, and let F be an event in Ω such that $\mathbf{P}(F) > 0$.

Definition 380 *For every $E \in \mathcal{E}$, we call the* conditional probability *of E given F the positive number given by*

$$\mathbf{P}(E \mid F) \overset{def}{=} \frac{\mathbf{P}(E \cap F)}{\mathbf{P}(F)}. \tag{4.107}$$

Loosely speaking, the conditional probability of the event E given F represents a re-evaluation of the probability of the occurrence of the conditioned event E in light of the occurrence of the conditioning event F. This is stressed by the following result

Proposition 381 *The map $\mathbf{P}_F : \mathcal{E} \to \mathbb{R}$ given by*

$$\mathbf{P}_F(E) \overset{def}{=} \mathbf{P}(E \mid F), \quad \forall E \in \mathcal{E}, \tag{4.108}$$

is a probability on Ω, which is concentrated *on F, to say $\mathbf{P}_F(F) = 1$ and $\mathbf{P}_F(E) = 0$ for every $E \in \mathcal{E}$ such that $E \cap F = \varnothing$.*

Proof. We clearly have

$$\mathbf{P}_F(F) = \frac{\mathbf{P}(F \cap F)}{\mathbf{P}(F)} = \frac{\mathbf{P}(F)}{\mathbf{P}(F)} = 1$$

and

$$\mathbf{P}_F(E) = \frac{\mathbf{P}(E \cap F)}{\mathbf{P}(F)} = \frac{\mathbf{P}(\varnothing)}{\mathbf{P}(F)} = 0,$$

for every $E \in \mathcal{E}$ such that $E \cap F = \varnothing$. Furthermore,

$$\mathbf{P}_F\left(\bigcup_{n=1}^{\infty} E_n\right) = \frac{\mathbf{P}\left(\left(\bigcup_{n=1}^{\infty} E_n\right) \cap F\right)}{\mathbf{P}(F)} = \frac{\mathbf{P}\left(\bigcup_{n=1}^{\infty} (E_n \cap F)\right)}{\mathbf{P}(F)}, \tag{4.109}$$

for every sequence $(E_n)_{n=1}^{\infty}$ in $\mathcal{E}$. Now, if the events of the sequence $(E_n)_{n=1}^{\infty}$ are pairwise incompatible, a fortiori the events of the sequence $(E_n \cap F)_{n=1}^{\infty}$ are pairwise incompatible. Hence,

$$\frac{\mathbf{P}\left(\bigcup_{n=1}^{\infty} (E_n \cap F)\right)}{\mathbf{P}(F)} = \frac{\sum_{n=1}^{\infty} \mathbf{P}(E_n \cap F)}{\mathbf{P}(F)} = \sum_{n=1}^{\infty} \frac{\mathbf{P}(E_n \cap F)}{\mathbf{P}(F)} = \sum_{n=1}^{\infty} \mathbf{P}_F(E_n). \tag{4.110}$$

Combining (4.109) and (4.110) we obtain the denumerable additivity of $\mathbf{P}_F : \mathcal{E} \to \mathbb{R}$. □

Definition 382 *We call the map* $\mathbf{P}_F : \mathcal{E} \to \mathbb{R}$ *introduced in Proposition 381 the* conditional probability *on* Ω given *the event* F*.*

Proposition 383 *We have*

$$\mathbf{P}(E \mid F) = \mathbf{P}(E), \tag{4.111}$$

if and only if the events E *and* F *are independent.*

Proof. If (4.111) holds true, we have

$$\frac{\mathbf{P}(E \cap F)}{\mathbf{P}(F)} = \mathbf{P}(E).$$

This clearly implies that Equation (4.100) holds true, that is the events E and F are independent. Conversely, if the events E and F are independent we have

$$\mathbf{P}(E \cap F) = \mathbf{P}(E)\,\mathbf{P}(F).$$

Combining the latter with Equation (4.107) it follows that (4.111) holds true. □

Example 384 *Two fair dice are rolled: a white die and a black one, denoted by* W *and* B*, respectively. Let us compute the naive probability of obtaining a double six, given that the white die shows a six. Let us also compute the naive probability of obtaining a double six, given that both the outcomes of the roll are even.*

Discussion. Let us represent the sample space of all possible outcomes of the roll of the die by the set $\Omega_k \equiv \{1, 2, \ldots, 6\}$, for $k = W, B$. Then, the sample space of all possible outcomes of the roll of the two dice is represented by the set $\Omega = \Omega_W \times \Omega_B \equiv \{(1,1), (1,2), \ldots, (6,6)\}$. The event "the white die shows a six" is Then, represented by the set $E_6 \times \Omega_B \equiv \{(6,1), (6,2), \ldots, (6,6)\}$, where $E_6 \equiv \{6\}$, and the event "a double six is obtained" is represented by the set $E_{6,6} \equiv \{(6,6)\}$. Therefore,

$$\mathbf{P}(E_{6,6} \mid E_6 \times \Omega_B) = \frac{\mathbf{P}(E_{6,6} \cap E_6 \times \Omega_B)}{\mathbf{P}(E_6 \times \Omega_B)} = \frac{\mathbf{P}(E_{6,6})}{\mathbf{P}(E_6 \times \Omega_B)} = \frac{\frac{|E_{6,6}|}{|\Omega|}}{\frac{|E_6 \times \Omega_B|}{|\Omega|}} = \frac{|E_{6,6}|}{|E_6 \times \Omega_B|} = \frac{1}{6}.$$

The event "both the outcomes of the roll are even" is represented by $E_{\mathbb{E},\mathbb{E}} \equiv \{(2,2), (2,4), \ldots, (6,6)\}$. Hence,

$$\mathbf{P}(E_{6,6} \mid E_{\mathbb{E},\mathbb{E}}) = \frac{\mathbf{P}(E_{6,6} \cap E_{\mathbb{E},\mathbb{E}})}{\mathbf{P}(E_{\mathbb{E},\mathbb{E}})} = \frac{\mathbf{P}(E_{6,6})}{\mathbf{P}(E_{\mathbb{E},\mathbb{E}})} = \frac{\frac{|E_{6,6}|}{|\Omega|}}{\frac{|E_{\mathbb{E},\mathbb{E}}|}{|\Omega|}} = \frac{|E_{6,6}|}{|E_{\mathbb{E},\mathbb{E}}|} = \frac{1}{9}.$$

Note that in both the cases we are considering a re-evaluation of the probability of the occurrence of the conditioned event in light of the occurrence of conditioning event. Note also that neither the events $E_{6,6}$ and $E_6 \times \Omega_B$ nor the events $E_{6,6}$ and $E_{\mathbb{E},\mathbb{E}}$ are independent. □

Example 385 *With reference to Example 301, assume that two balls are drawn in sequence from the urn without replacement. Let us compute the naive probability that the second ball drawn is white, given that the first one is white and given that the first one is black. Let us also compute the naive probability that the second ball drawn is white if we ignore the colour of the first one.*

Discussion. We already discussed this issue by exploiting the bivariate hypergeometric distribution (see Example 307). Here, we consider an approach based on conditional probability. Write W_k for the event "the kth ball drawn is white" for $k = 1, 2$ and B_1 for the event "the first ball drawn is black". We know that

$$\mathbf{P}(W_1) = \frac{M}{N} \quad \text{and} \quad \mathbf{P}(B_1) = \frac{N-M}{N}.$$

We want to compute

$$\mathbf{P}(W_2 \mid W_1) \quad \text{and} \quad \mathbf{P}(W_2 \mid B_1).$$

Under our assumption about the composition of the urn and the drawning technique, given that a white [resp. black] ball is drawn at the first draw, $N-1$ balls of which $M-1$ [resp. M] are white remain in the urn. Therefore, the naive probability of drawing a white ball at the second draw given that a white [resp. black] ball is drawn at the first draw, is

$$\mathbf{P}(W_2 \mid W_1) = \frac{M-1}{N-1} \qquad [\text{resp. } \mathbf{P}(W_2 \mid B_1) = \frac{M}{N-1}].$$

In case we ignore the colour of the first ball drawn, we have to compute $\mathbf{P}(W_2)$. However, we know for sure that the first ball drawn is white or black, that is

$$W_1 \cup B_1 = \Omega.$$

Therefore, we can compute

$$\begin{aligned} \mathbf{P}(W_2) &= \mathbf{P}(W_2 \cap \Omega) = \mathbf{P}(W_2 \cap (W_1 \cup B_1)) \\ &= \mathbf{P}(W_2 \cap W_1) + \mathbf{P}(W_2 \cap B_1) \\ &= \mathbf{P}(W_2 \mid W_1)\,\mathbf{P}(W_1) + \mathbf{P}(W_2 \mid B_1)\,\mathbf{P}(B_1) \\ &= \frac{M-1}{N-1} \cdot \frac{M}{N} + \frac{M}{N-1} \cdot \frac{N-M}{N} \\ &= \frac{M}{N}. \end{aligned} \tag{4.112}$$

This result, somewhat counterintuitive since $\mathbf{P}(W_2) = \mathbf{P}(W_1)$, as it were if the balls were drawn with replacement, might be better grasped considering that after the first unobserved draw the composition of the urn has changed in a random way. More precisely, after the first unobserved draw, we have an urn with $M-1$ [resp. M] white balls with probability $\mathbf{P}(W_1)$ [resp. $\mathbf{P}(B_1)$]. Note that the fourth line of Equation (4.112) shows a favorable/possible cases scheme weighted by the occurrence probabilities. In the end, accounting for the random change of the urn is the same as ignoring the change. □

Proposition 386 (simmetry formula) *Let E, F be events in Ω such that $\mathbf{P}(E), \mathbf{P}(F) > 0$. Then,*

$$\frac{\mathbf{P}(E \mid F)}{\mathbf{P}(E)} = \frac{\mathbf{P}(F \mid E)}{\mathbf{P}(F)}. \tag{4.113}$$

Proof. By definition,

$$\mathbf{P}(E \mid F)\,\mathbf{P}(F) = \mathbf{P}(E \cap F) \quad \text{and} \quad \mathbf{P}(F \mid E)\,\mathbf{P}(E) = \mathbf{P}(F \cap E).$$

Equation (4.113) immediately follows. □

Theorem 387 (total probability formula) *Let $N \subseteq \mathbb{N}$ and let $(F_n)_{n\in N}$ be a partition of Ω such that $\mathbf{P}(F_n) > 0$, for every $n \in N$. We have*

$$\mathbf{P}(E) = \sum_{n\in N} \mathbf{P}(E \mid F_n)\,\mathbf{P}(F_n), \tag{4.114}$$

for every $E \in \mathcal{E}$.

Proof. We can write

$$\mathbf{P}(E) = \mathbf{P}(E \cap \Omega) = \mathbf{P}\left(E \cap \left(\bigcup_{n\in N} F_n\right)\right) = \mathbf{P}\left(\bigcup_{n\in N} (E \cap F_n)\right) = \sum_{n\in N} \mathbf{P}(E \cap F_n). \tag{4.115}$$

On the other hand, by definition

$$\mathbf{P}(E \cap F_n) = \mathbf{P}(E \mid F_n)\,\mathbf{P}(F_n), \tag{4.116}$$

for every $n \geq 1$. Combining (4.115) and (4.116), the desired (4.114) immediately follows. □

Theorem 388 (Bayes formula) *Let $N \subseteq \mathbb{N}$ and let $(F_n)_{n\in N}$ be a partition of Ω such that $\mathbf{P}(F_n) > 0$, for every $n \in N$. Then, we have*

$$\mathbf{P}(F_n \mid E) = \frac{\mathbf{P}(E \mid F_n)\,\mathbf{P}(F_n)}{\sum_{m\in N} \mathbf{P}(E \mid F_m)\,\mathbf{P}(F_m)}, \tag{4.117}$$

for every $n \in N$ and every $E \in \mathcal{E}$ such that $\mathbf{P}(E) > 0$.

Proof. From the symmetry formula (4.113), we can write

$$\mathbf{P}(F_n \mid E) = \frac{\mathbf{P}(E \mid F_n)\,\mathbf{P}(F_n)}{\mathbf{P}(E)}, \tag{4.118}$$

for any $n \in N$. Hence, combining (4.114) and (4.118), we obtain Equation (4.117). □

Example 389 *Consider an urn containing $N \geq 2$ balls of which M are white and $N - M$ are black, for $1 \leq M < N$. Assume the balls are drawn from the urn with replacement. Then, the naive probability of drawing a white ball at the nth drawn is M/N and the naive probability of drawing the white ball at the nth draw, given that all previous draws have failed is still M/N.*

Discussion. Under the assumption that the balls are drawn with replacement, at the nth draw the urn has the same composition it has at the first draw. Therefore, to deal with a sequence of draws we can think on a sequence of urns all with the same composition. If we number the ball in each urn from 1 to N while numbering the white balls from 1 to M, we can clearly represent the sample space by the cartesian product

$$\Omega \equiv \mathsf{X}_{k=1}^{n}\Omega_k, \quad \Omega_k \equiv \{1, \ldots, N\}, \quad k = 1, \ldots, n.$$

Writing W_n for the event "a white ball is drawn at the nth draw" it is clear that W_n is represented by the set on all n-tuples of Ω showing a number in $\{1, \ldots, M\}$ at the nth entry, that is $W_n \equiv \mathsf{X}_{k=1}^{n-1}\Omega_k \times E_n$, where $E_n \equiv \{1, \ldots, M\}$. Then, the naive probability of W_n is

$$\mathbf{P}(W_n) = \frac{\left|\mathsf{X}_{k=1}^{n-1}\Omega_k \times E_n\right|}{|\Omega|} = \frac{\Pi_{k=1}^{n-1}|\Omega_k| \cdot |E_n|}{\Pi_{k=1}^{n}|\Omega_k|} = \frac{|E_n|}{|\Omega_k|} = \frac{M}{N}.$$

Write B_k for the event "a black ball is drawn at the kth draw", with $1 \leq k < n$. We can represent B_k by the set on all n-tuples of Ω showing a number in $\{M+1, \ldots, N\}$ at the kth entry, that is $B_k \equiv \mathsf{X}_{j=1}^{k-1}\Omega_j \times F_k \times \mathsf{X}_{j=k+1}^{n}\Omega_j$, where $F_k \equiv \{M+1, \ldots, N\}$. The naive probability of B_k turns out to be

$$\mathbf{P}(B_k) = \frac{\left|\mathsf{X}_{j=1}^{k-1}\Omega_j \times F_k \times \mathsf{X}_{j=k+1}^{n}\Omega_j\right|}{|\Omega|} = \frac{\Pi_{j=1}^{k-1}|\Omega_k| \cdot |F_n| \cdot \Pi_{j=k+1}^{n}|\Omega_k|}{\Pi_{k=1}^{n}|\Omega_k|} = \frac{|F_n|}{|\Omega_k|} = \frac{N-M}{N}.$$

In addition, writing $B_{1,\ldots,n-1}$ for the event "a black ball is drawn at the kth draw for $k = 1, \ldots, n-1$", we can represent $B_{1,\ldots,n-1}$ by the set on all n-tuples of Ω showing a number in $\{M+1, \ldots, N\}$ at the kth entry, for $k = 1, \ldots, n-1$. We Then, have

$$\mathbf{P}(B_{1,\ldots,n-1}) = \frac{\left|\mathsf{X}_{k=1}^{n-1}F_k \times \Omega_n\right|}{|\Omega|} = \frac{\Pi_{k=1}^{n-1}|F_k| \cdot |\Omega_n|}{\Pi_{k=1}^{n}|\Omega_k|} = \frac{(N-M)^{n-1}}{N^{n-1}} = \left(\frac{N-M}{N}\right)^{n-1}.$$

Now, $B_{1,\ldots,n-1} \cap W_n$ is the event "a black ball is drawn at the kth draw, for $k = 1, \ldots, n-1$, and a white ball is drawn at the nth draw". We can represent $B_{1,\ldots,n-1} \cap W_n$ by the set on all n-tuples of Ω showing a number in $\{M+1, \ldots, N\}$ at the kth entry, for $k = 1, \ldots, n-1$ and a number in $\{1, \ldots, M\}$ at the nth entry. It follows

$$\mathbf{P}(B_{1,\ldots,n-1} \cap W_n) = \frac{\left|\mathsf{X}_{k=1}^{n-1}F_k \times E_n\right|}{|\Omega|} = \frac{\Pi_{k=1}^{n-1}|F_k| \cdot |E_n|}{\Pi_{k=1}^{n}|\Omega_k|} = \frac{(N-M)^{n-1}M}{N^n} = \left(\frac{N-M}{N}\right)^{n-1}\frac{N}{N}.$$

In light of what shown above, the probability of the event "a white ball is drawn at the nth draw, given that all previous draws have failed" is given by

$$\mathbf{P}(W_n \mid B_{1,\ldots,n-1}) = \frac{\mathbf{P}(B_{1,\ldots,n-1} \cap W_n)}{\mathbf{P}(B_{1,\ldots,n-1})} = \frac{\left(\frac{N-M}{N}\right)^{n-1}\frac{N}{N}}{\left(\frac{N-M}{N}\right)^{n-1}} = \frac{N}{N}.$$

Note that

$$\mathbf{P}(W_n \mid B_{1,\ldots,n-1}) = \mathbf{P}(W_n).$$

Hence, in the model considered, the events W_n and $B_{1,\ldots,n-1}$ are independent. □

Example 390 *Consider an urn containing $N \geq 2$ balls of which M are white and $N-M$ are black, for $1 \leq M < N$. Assume the balls are drawn from the urn* without replacement. *Then, the naive probability of drawing a white ball at the nth draw, given that all previous draws have failed, is $M/(N-n+1)$, for every $n = 1, \ldots, N-M+1$, while the naive probability of drawing a white ball at the nth draw, for every $n = 1, \ldots, N$, is still M/N.*

Discussion. As in Example 389, we number the balls in the urn from 1 to N while numbering the white balls from 1 to M. Note that the maximum number of black balls that can be drawn from the urn is $N-M$. Therefore, it is possible to draw a white ball at the nth draw, given that all previous draws have failed, provided that $n-1 \leq M-N$, that is $n \leq N-M+1$. In particular, when $n = N-M+1$ we obtain $M/(N-n+1) = 1$, as it should be. Now, for any $n \leq N-M+1$, write $W_{n-1,k}^{N,M}$ for the event "from an urn containing N white balls of which $M \leq N$ are white, k white balls are drawn in $n-1$ draws", for $k = 0, 1, \ldots, n-1$. The family

$\left(W_{n-1,k}^{N,M}\right)_{k=0}^{n-1}$ is a partition of the sure event. In addition, applying the bivariate hypergeometric distribution, we know that

$$\mathbf{P}\left(W_{n-1,k}^{N,M}\right) = \frac{\binom{M}{k}\binom{N-M}{n-1-k}}{\binom{N}{n-1}}. \tag{4.119}$$

Note that, since we clearly have

$$\sum_{k=0}^{n-1}\mathbf{P}\left(W_{n-1,k}^{N,M}\right) = 1,$$

we obtain a probabilistic proof of the Vandermonde identity (see Equation (4.26)). Write $W_n^{N,M}$ for the event "from an urn containing N white balls of which $M \leq N$ are white, the ball drawn at nth draw is white". By virtue of the total probability formula, we have

$$\mathbf{P}\left(W_n^{N,M}\right) = \sum_{k=0}^{n-1}\mathbf{P}\left(W_n^{N,M} \mid W_{n-1,k}^{N,M}\right)\mathbf{P}\left(W_{n-1,k}^{N,M}\right).$$

On the other hand, applying the naive probability, we clearly have

$$\mathbf{P}\left(W_n^{N,M} \mid W_{n-1,k}^{N,M}\right) = \frac{M-k}{N-(n-1)},$$

for every $k = 0, \ldots, n-1$. In particular, probability of drawing a white ball at the nth draw, given that all previous drawns have failed is

$$\mathbf{P}\left(W_n^{N,M} \mid W_{n-1,0}^{N,M}\right) = \frac{M}{N-(n-1)}.$$

Now,

$$\begin{aligned}
\mathbf{P}\left(W_n^{N,M}\right) &= \sum_{k=0}^{n-1}\mathbf{P}\left(W_n^{N,M} \mid W_{n-1,k}^{N,M}\right)\mathbf{P}\left(W_{n-1,k}^{N,M}\right) \\
&= \sum_{k=0}^{n-1}\frac{M-k}{N-(n-1)}\frac{\binom{M}{k}\binom{N-M}{n-1-k}}{\binom{N}{n-1}} \\
&= \sum_{k=0}^{n-1}\frac{M-k}{N-(n-1)}\frac{M!}{(M-k)!k!}\binom{N-1-(M-1)}{n-1-k}\frac{(n-1)!\,(N-(n-1))!}{N!} \\
&= \frac{M}{N}\sum_{k=0}^{n-1}\frac{(M-1)!}{(M-1-k)!k!}\binom{N-1-(M-1)}{n-1-k}\frac{(n-1)!\,(N-1-(n-1))!}{(N-1)!} \\
&= \frac{M}{N}\sum_{k=0}^{n-1}\frac{\binom{M-1}{k}\binom{N-1-(M-1)}{n-1-k}}{\binom{N-1}{n-1}}.
\end{aligned} \tag{4.120}$$

On the other hand, writing $W_{n-1,k}^{N-1,M-1}$ for the event "from an urn containing $N-1$ balls of which $M-1 \leq N-1$ are white, k white balls are drawn in $n-1$ draws", for $k=0,1,\ldots,n-1$, and comparing to (4.119), we have

$$\mathbf{P}\left(W_{n-1,k}^{(N-1,M-1)}\right) = \frac{\binom{M-1}{k}\binom{N-1-(M-1)}{n-1-k}}{\binom{N-1}{n-1}}. \tag{4.121}$$

As a consequence

$$\sum_{k=0}^{n-1} \frac{\binom{M-1}{k}\binom{N-1-(M-1)}{n-1-k}}{\binom{N-1}{n-1}} = \sum_{k=0}^{n-1} \mathbf{P}\left(W_{n-1,k}^{(N-1,M-1)}\right) = 1. \tag{4.122}$$

Combining (4.120)-(4.122), it Then, follows,

$$\mathbf{P}\left(W_n^{N,M}\right) = \frac{M}{N}.$$

as desired. □

In applications, the main difficulty in dealing with conditioned probability is the identification of the events whose probabilities and conditioned probabilities we aim to compute. This issue can be exemplified by the following problems.

Problem 391 *The National Health Service (NHS) aims to introduce a new test for the screening of a disease. The pharmaceutical company which produces the test states that:*

- *the test yields a positive result on the* 95% *of people who are affected by the disease (*sensitivity *or* true positive rate *of the test);*
- *the test yields a negative result on the* 99% *of people who are not affected by the disease (*specificity *or* true negative rate *of the test);*

On the other hand, the NHS knows the the disease is currently affecting the 10% *of the population. Compute:*

1. *the probability that a randomly chosen individual of the population is affected by the disease given that the test yields a positive result;*
2. *the probability that a randomly chosen individual of the population is not affected by the disease given that the test yields a positive result;*
3. *the probability that a randomly chosen individual of the population is affected by the disease given that the test yields a negative result;*
4. *the probability that a randomly chosen individual of the population is not affected by the disease given that the test yields a negative result;*

5. *the probability that the test yields a positive result on a randomly chosen individual of the population.*

6. *the probability that the test yields a negative result on a randomly chosen individual of the population.*

Solution. Write D [resp. H] for the event "a randomly chosen individual of the population is affected [resp. not affected] by the disease". We have

$$D \cup H = \Omega \quad \text{and} \quad D \cap H = \varnothing.$$

It follows

$$\mathbf{P}(D) + \mathbf{P}(H) = 1. \tag{4.123}$$

Write T_+ [resp. T_-] for the event "the test yields a positive [resp. negative] result on a randomly chosen individual of the population". We have

$$T_+ \cup T_- = \Omega \quad \text{and} \quad T_+ \cap T_- = \varnothing.$$

It follows

$$\mathbf{P}(T_+) + \mathbf{P}(T_-) = 1. \tag{4.124}$$

In terms of the above notations, the information provided by the pharmaceutical company means

$$\mathbf{P}(T_+ \mid D) = 0.95, \qquad \text{and} \qquad \mathbf{P}(T_- \mid H) = 0.99. \tag{4.125}$$

The information provided by NHS is

$$\mathbf{P}(D) = 0.10, \tag{4.126}$$

which clearly implies

$$\mathbf{P}(H) = 0.90. \tag{4.127}$$

To answer Questions **??-??**, we need to compute the following probabilities

$$\mathbf{P}(D \mid T_+), \quad \mathbf{P}(H \mid T_+), \quad \mathbf{P}(D \mid T_-), \quad \mathbf{P}(H \mid T_-), \quad \mathbf{P}(T_+), \quad \mathbf{P}(T_-),$$

respectively. Now, from the symmetry formula of conditional probabilities and on account of (4.125)-(4.127), we know that

$$\mathbf{P}(D \mid T_+) = \frac{\mathbf{P}(T_+ \mid D)\,\mathbf{P}(D)}{\mathbf{P}(T_+)} = \frac{0.95 * 0.10}{\mathbf{P}(T_+)} = \frac{0.095}{\mathbf{P}(T_+)}, \tag{4.128}$$

$$\mathbf{P}(H \mid T_+) = \frac{\mathbf{P}(T_+ \mid H)\,\mathbf{P}(H)}{\mathbf{P}(T_+)} = \frac{0.90 * \mathbf{P}(T_+ \mid H)}{\mathbf{P}(T_+)}, \tag{4.129}$$

$$\mathbf{P}(D \mid T_-) = \frac{\mathbf{P}(T_- \mid D)\,\mathbf{P}(D)}{\mathbf{P}(T_-)} = \frac{0.10 * \mathbf{P}(T_- \mid D)}{\mathbf{P}(T_-)}, \tag{4.130}$$

$$\mathbf{P}(H \mid T_-) = \frac{\mathbf{P}(T_- \mid H)\,\mathbf{P}(H)}{\mathbf{P}(T_-)} = \frac{0.99 * 0.90}{\mathbf{P}(T_-)} = \frac{0.891}{\mathbf{P}(T_-)}, \tag{4.131}$$

On the other hand, we know that the conditional probability is a probability concentrated on the conditioning event. Hence, we have

$$\mathbf{P}(T_+ \mid H) = 1 - \mathbf{P}(T_- \mid H) = 1 - 0.99 = 0.01. \tag{4.132}$$

and

$$\mathbf{P}\left(T_{-} \mid D\right) = 1 - \mathbf{P}\left(T_{+} \mid D\right) = 1 - 0.95 = 0.05 \tag{4.133}$$

Therefore, on account of (4.124), to answer Questions **??-??**, we are left with computing $\mathbf{P}\left(T_{+}\right)$. To this, by the Total Probability Formula, we can write

$$\begin{aligned}\mathbf{P}\left(T_{+}\right) &= \mathbf{P}\left(T_{+} \mid H\right)\mathbf{P}\left(H\right) + \mathbf{P}\left(T_{+} \mid D\right)\mathbf{P}\left(D\right) \\ &= 0.01 * 0.90 + 0.95 * 0.10 = 0.104. \end{aligned} \tag{4.134}$$

As a consequence,

$$\mathbf{P}\left(T_{-}\right) = 1 - 0.104 = 0.896. \tag{4.135}$$

In the end, replacing (4.132)-(4.135) into (4.128)-(4.131), we obtain

$$\begin{aligned}\mathbf{P}\left(D \mid T_{+}\right) &= \frac{0.095}{0.104} = 0.913, \\ \mathbf{P}\left(H \mid T_{+}\right) &= \frac{0.90 * 0.01}{0.104} = 8.654 \times 10^{-2}, \\ \mathbf{P}\left(D \mid T_{-}\right) &= \frac{0.10 * 0.05}{0.896} = 5.58 \times 10^{-3}, \\ \mathbf{P}\left(H \mid T_{-}\right) &= \frac{0.891}{0.896} = 0.994,\end{aligned}$$

which complete the answers. □

Problem 392 (Monty's hall) *A quiz master shows a participant three closed boxes labeled by* A, B, C*. One of the boxes, chosen by the quiz organisers with uniform probability, contains a prize of* $\$1,000$*. The remaining two are empty. The quiz master asks the participant to choose a box. Once the participant makes own choice the quiz master opens one of the two rejected boxes and shows that it is empty. Thereafter, the quiz master gives the participant the opportunity either to stick to the first choice or to exchange the chosen box with the one of the rejected boxes which is still closed. Assume the quiz master knows what box contains the prize, the quiz master never shows a box containing the prize, and the quiz master chooses a empty box between two with uniform probability. What should the participant do? To stick to her first choice, to accept the exchange or it does not matter at all because the odds are now fifty-fifty?*

Solution. Assume the quiz participant chooses box A and thereafter the quiz master shows the empty box B or C. Denote by PX the event "the prize is in box X" for $X = A, B, C$ and EY the event "the quiz master shows empty box Y", for $Y = B, C$. Under the assumptions of the problem, we have

$$\mathbf{P}\left(PA\right) = \mathbf{P}\left(PB\right) = \mathbf{P}\left(PC\right) = \frac{1}{3}.$$

Now, since the family of events $\{PA, PB, PC\}$ is a partition of the sure event Ω, by virtue of the Total Probability Formula (4.114), we have

$$\mathbf{P}\left(EB\right) = \mathbf{P}\left(EB \mid PA\right)\mathbf{P}\left(PA\right) + \mathbf{P}\left(EB \mid PB\right)\mathbf{P}\left(PB\right) + \mathbf{P}\left(EB \mid PC\right)\mathbf{P}\left(PC\right) \tag{4.136}$$

and

$$\mathbf{P}\left(EC\right) = \mathbf{P}\left(EC \mid PA\right)\mathbf{P}\left(PA\right) + \mathbf{P}\left(EC \mid PB\right)\mathbf{P}\left(PB\right) + \mathbf{P}\left(EC \mid PC\right)\mathbf{P}\left(PC\right) \tag{4.137}$$

Furthermore, the assumptions of the problem still entail

$$\begin{array}{lll} \mathbf{P}(EB \mid PA) = \frac{1}{2}, & \mathbf{P}(EB \mid PB) = 0, & \mathbf{P}(EB \mid PC) = 1, \\ \mathbf{P}(EC \mid PA) = \frac{1}{2}, & \mathbf{P}(EC \mid PB) = 1, & \mathbf{P}(EC \mid PC) = 0. \end{array} \tag{4.138}$$

Combining (4.136), (4.137) and (4.138), we obtain

$$\mathbf{P}(EB) = \mathbf{P}(EC) = \frac{1}{2}. \tag{4.139}$$

Note that this is not the same as the assumption

$$\mathbf{P}(EB \mid PA) = \mathbf{P}(EC \mid PA) = \frac{1}{2}.$$

By (4.113), it Then, follows

$$\mathbf{P}(PA \mid EB) = \frac{\mathbf{P}(EB|PA)\mathbf{P}(PA)}{\mathbf{P}(EB)} = \frac{\frac{1}{2}\cdot\frac{1}{3}}{\frac{1}{2}} = \frac{1}{3}, \quad \mathbf{P}(PA \mid EC) = \frac{\mathbf{P}(EC|PA)\mathbf{P}(PA)}{\mathbf{P}(EC)} = \frac{\frac{1}{2}\cdot\frac{1}{3}}{\frac{1}{2}} = \frac{1}{3}, \tag{4.140}$$

and

$$\mathbf{P}(PC \mid EB) = \frac{\mathbf{P}(EB|PC)\mathbf{P}(PC)}{\mathbf{P}(EB)} = \frac{1\cdot\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}, \quad \mathbf{P}(PB \mid EC) = \frac{\mathbf{P}(EC|PB)\mathbf{P}(PB)}{\mathbf{P}(EC)} = \frac{1\cdot\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}. \tag{4.141}$$

Hence, whether the quiz master shows empty box B or C the conditional probability that the prize is in the other box C or B is higher than the conditional probability that the prize is contained in the initially chosen box A. We can conclude that the quiz participant should exchange the initially chosen box A with closed box C or B.

Another argument is the following. Assume the participant chooses box A. Denote by V the event "the participant wins", by PA the event "the prize is in box A" and by PA^c the event "the prize is not in box A". We have clearly

$$\mathbf{P}(PA) = \frac{1}{3}, \quad \mathbf{P}(PA^c) = \frac{2}{3}.$$

The total probability formula (4.114) yields

$$\mathbf{P}(V) = \mathbf{P}(V \mid PA)\,\mathbf{P}(PA) + \mathbf{P}(V \mid PA^c)\,\mathbf{P}(PA^c) = \frac{1}{3}\mathbf{P}(V \mid PA) + \frac{2}{3}\mathbf{P}(V \mid PA^c). \tag{4.142}$$

Now, if the participant sticks to the choice of box A we have

$$\mathbf{P}(V \mid PA) = 1, \quad \mathbf{P}(V \mid PA^c) = 0,$$

which implies

$$\mathbf{P}(V) = \frac{1}{3}.$$

In contrast, if the participant exchanges the chosen box with the still-closed rejected box we have

$$\mathbf{P}(V \mid PA) = 0, \quad \mathbf{P}(V \mid PA^c) = 1,$$

which implies

$$\mathbf{P}(V) = \frac{2}{3}.$$

This confirms that the participant should exchange chosen box A with closed box B or C. □

Chapter 5

Real Random Variables

As discussed above, a probability space is a mathematical model to represent a random phenomenon. We now introduce another essential tool, termed *random variable*, to represent a quantitative or categorical observation of random phenomena. Such observations may return different values depending on different outcomes of the random phenomenon under concern. Hence, it is natural to represent them as functions on the probability space. On the other hand, the observation result must be detectable by an observer in light of her information: the observer must be able to discriminate whether the result of her observation takes a specific value or falls within a particular set of values. Therefore, what makes an ordinary function a random variable is the possibility of observing the events which allow discriminating among the possible values taken by the function, in light of the available information.

5.1 Complete Kolmogorov Probability Spaces

Let Ω be a sample space, let $\mathcal{E}$ be a σ-algebra of events of Ω, and let $\mathbf{P} : \mathcal{E} \to \mathbb{R}_+$ be a probability, denoted briefly by $\mathbf{P}$ when no confusion can arise about the domain $\mathcal{E}$.

Definition 393 (probability space) *We call the triple* $(\Omega, \mathcal{E}, \mathbf{P})$ *a* (Kolmogorov) probability space.

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ a probability space. We recall that the event Ω [resp. $\varnothing$] is referred to as the *sure* [resp. *impossible*] *event.*

Definition 394 *Given an event* $E \in \mathcal{E}$*, we say that* E *is* almost sure *if* $\mathbf{P}(E) = 1$.

Definition 395 *Given an event* $E \subseteq \Omega$*, we say that* E *is* negligible *if there exists an event* $F \in \mathcal{E}$ *such that* $E \subseteq F$ *and* $\mathbf{P}(F) = 0$.

Definition 396 *We say that a probability space* $(\Omega, \mathcal{E}, \mathbf{P})$ *is* complete *if* $\mathcal{E}$ *contains the family of all negligible events.*

Theorem 397 *Given any probability space* $(\Omega, \mathcal{E}, \mathbf{P})$*, there aways exists a unique complete probability space* $(\Omega, \overline{\mathcal{E}}, \overline{\mathbf{P}})$ *such that*

$$\mathcal{E} \subseteq \overline{\mathcal{E}} \quad \textit{and} \quad \overline{\mathbf{P}}_{|\mathcal{E}} = \mathbf{P}. \tag{5.1}$$

More specifically, writing $\mathcal{N}$ for the family of all negligible events in $(\Omega, \mathcal{E}, \mathbf{P})$, we have

$$\overline{\mathcal{E}} = \sigma(\mathcal{E}, \mathcal{N}) \quad \text{and} \quad \overline{\mathbf{P}}(N \cup E) = \mathbf{P}(E), \tag{5.2}$$

for any $N \in \mathcal{N}$ and $E \in \mathcal{E}$.

Proof. See e.g. Rick Durret, *Probability: Theory and Examples*, third edition, Thomson, Brooks/Cole, 2013, p. 450. See also Patrick Billingsley, *Probability and Measure*, third edition, Johm Wiley and Sons, 1995, Problems 3.10 and 10.5. □

Definition 398 (complete probability space) *The probability space $(\Omega, \overline{\mathcal{E}}, \overline{\mathbf{P}})$ introduced in Theorem 397, is called the* completion *of $(\Omega, \mathcal{E}, \mathbf{P})$.*

In light of Theorem 397, we can always deal with complete probability spaces. Therefore, from now on, by a probability space we mean a complete probability space.

5.2 Real Random Variables

Let $(\Omega, \mathcal{E}, \mathbf{P})$ be a probability space, let $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be the Euclidean real line equipped with the Borel σ-algebra $\mathcal{B}(\mathbb{R})$, and let $X : \Omega \to \mathbb{R}$, briefly X, be a real function on Ω.

Definition 399 (state space) *Following the standard probabilistic terminology, we call the points $x \in \mathbb{R}$* states *and we call the couple $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ the* real Borel state space.

Notation 400 *For brevity, it is customary to denote the probability space $(\Omega, \mathcal{E}, \mathbf{P})$ with the single symbol Ω rather that the triple $(\Omega, \mathcal{E}, \mathbf{P})$, when no confusion can arise about the σ-algebra of events $\mathcal{E}$ or the probability $\mathbf{P}$. Similarly, we use the single symbol $\mathbb{R}$ rather that the couple $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ to denote the real Borel state space.*

Notation 401 *Given any $B \in \mathcal{B}(\mathbb{R})$, we will use the standard probability shorthand $\{X \in B\}$, rather than the set-theory notation $X^{-1}(B)$, as an abbreviation to represent the event $\{\omega \in \Omega : X(\omega) \in B\}$, referred to as the X-inverse image of B.*

Definition 402 *We say that X is a* real $\mathcal{E}$-random variable*, if the X-inverse image of any Borel set in $\mathcal{B}(\mathbb{R})$ is an event in $\mathcal{E}$. In symbols,*

$$\{X \in B\} \in \mathcal{E}, \quad \forall B \in \mathcal{B}(\mathbb{R}), \tag{5.3}$$

In case the σ-algebra of events $\mathcal{E}$ on Ω is fixed once and for all, so that there is no danger of confusion, we will omit mentioning it and speak simply of real random variable*.*

Definition 403 *If X is is a* real $\mathcal{E}$-random variable*, then, for any $\omega \in \Omega$ the value $X(\omega) \in \mathbb{R}$, referred to as the X-image of ω, is also called a* state *or a* realization *of X.*

We already mentioned that the available information on a random phenomenon is represented by the σ-algebra $\mathcal{E}$ of the events on the sample space Ω. Therefore, Equation (5.3) is well suited to represent the idea of observability of the discriminating events. Note also that the notion of random variable on a probabililty space with values in a state space corresponds faithfully to the notion of *measurable map* from a *measure space* to a *measurable space*. As

a consequence, all fundamental results for measurable maps developed within *measure theory* remain valid for random variables.

Note that, in the context of measure theory or functional analysis, the triple $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ [resp. couple $(\mathbb{R}, \mathcal{B}(\mathbb{R})) \equiv \mathbb{R}$] is referred to as a *measure space* [resp. *measurable space*]. Furthermore, the sets $B \in \mathcal{B}(\mathbb{R})$ are referred to as *measurable sets.*

Example 404 (Dirac random variable) *Fixed any $x_0 \in \mathbb{R}$, the function $X : \Omega \to \mathbb{R}$ given by*

$$X(\omega) \overset{def}{=} x_0, \quad \forall \omega \in \Omega, \tag{5.4}$$

is a real $\mathcal{E}$-random variable whatever the σ-algebra $\mathcal{E}$ is.

Discussion. Consider any $B \in \mathcal{M}$. Only two cases are possible: $x_0 \in B$ or $x_0 \notin B$. In case $x_0 \in B$, we have

$$\{X \in B\} = \Omega \in \mathcal{E}.$$

In case $x_0 \notin B$, we have

$$\{X \in B\} = \varnothing \in \mathcal{E}.$$

Therefore, regardless of how $\mathcal{E}$ is assigned, X is always a real $\mathcal{E}$-random variable. □

Definition 405 *We call the random variable given by Equation (21.3) the* Dirac random variable *concentrated at x_0.*

Definition 406 (Dirac random variable) *We call the random variable defined by Equation (??) the* Dirac *real random variable concentrated at x_0 and denote it by Dirac(x_0). In case $x_0 \equiv 0$, we speak of the* standard Dirac *real random variable.*

A Dirac random variable is also referred to as a *deterministic* or *constant* random variable, in the sense that the value x_0 it takes does not change on varying of the sample point $\omega \in \Omega$. The Dirac random variable constitutes a model for a deterministic observation in a probabilistic setting.

Remark 407 (Dirac random variable) *A function $X : \Omega \to \mathbb{R}$ is a Dirac random variable if and only if we have*

$$X = x_0 1_\Omega, \tag{5.5}$$

for a point $x_0 \in \mathbb{R}$.

Example 408 *Assume $\mathcal{E}$ is the trivial information on Ω, that is $\mathcal{E} = \{\varnothing, \Omega\}$. Then, the only functions $X : \Omega \to \mathbb{R}$ which are real $\mathcal{E}$-random variables are the Dirac random variables.*

By contrast,

Example 409 *Assume $\mathcal{E}$ is the complete information on Ω, that is $\mathcal{E} = \mathcal{P}(\Omega)$. Then, any function $X : \Omega \to \mathbb{R}$ is a real $\mathcal{E}$-random variable.*

According to Definition 402 and Examples 408 and 409, it appears rather clear that the property of being a random variable depends on the σ.algebra $\mathcal{E}$ assigned on the sample space, that is, on the available information on the random phenomenon. By changing the available information, the set of random variables also changes. This is stressed by the following example.

Example 410 *Let $\Omega \equiv \{\omega_1, \ldots, \omega_6\}$ be the sample space of all possible outcomes of the roll of a fair die and let $X : \Omega \to \mathbb{R}$ the function given by*

$$X(\omega_k) \overset{def}{=} \begin{cases} k & \text{if } k \text{ is odd} \\ -k+1 & \text{if } k \text{ is even} \end{cases}, \quad \forall k = 1, \ldots, 6.$$

If $\mathcal{E} \equiv \mathcal{P}(\Omega)$ is the complete information on Ω, Then, X is clearly a real $\mathcal{E}$-random variable. On the other hand, if the availabble information is represented by the σ-algenbra of events $\mathcal{E} \equiv \{\varnothing, \Omega, \{\omega_1, \omega_3, \omega_5\}, \{\omega_2, \omega_4, \omega_6\}\}$, then, X is not a real $\mathcal{E}$-random variable.

Discussion. It is sufficient to observe that if $\mathcal{E} \equiv \{\varnothing, \Omega, \{\omega_1, \omega_3, \omega_5\}, \{\omega_2, \omega_4, \omega_6\}\}$, we have

$$(1/2, 3/2) \in \mathcal{B}(\mathbb{R}) \quad \text{and} \quad \{X \in (1/2, 3/2)\} = \{\omega_1\} \notin \mathcal{E}.$$

This implies that X is not a real $\mathcal{E}$-random variable. □

Remark 411 *Let $\mathcal{F}$ be a σ-algebra of events of Ω such that $\mathcal{E} \subseteq \mathcal{F}$. Then, any real $\mathcal{E}$-random variable is also an $\mathcal{F}$-random variable.*

We have

Proposition 412 *Given any function $X : \Omega \to \mathbb{R}$, the family of events of Ω*

$$\sigma(X) \overset{def}{=} \{E \in \mathcal{P}(\Omega) : E = \{X \in B\},\ B \in \mathcal{B}(\mathbb{R})\}, \tag{5.6}$$

is a σ-algebra

Proof. Clearly $\sigma(X)$ is not empty, since $\varnothing = \{X \in \varnothing\}$ and $\Omega = \{X \in \mathbb{R}\}$ are in $\sigma(X)$. Assume $E \in \sigma(X)$, Then, we have $E = \{X \in B\}$, for some $B \in \mathcal{B}(\mathbb{R})$. On the other hand, we know that

$$E^c \equiv \{X \in B\}^c = \{X \in B^c\},$$

where $B^c \in \mathcal{B}(\mathbb{R})$, since $B \in \mathcal{B}(\mathbb{R})$. It then follows that $E^c \in \sigma(X)$. Now, assume $(E_n)_{n \geq 1}$ is a sequence in $\sigma(X)$, then, for every $n \geq 1$ there exists $B_n \in \mathcal{B}(\mathbb{R})$ such that $E_n = \{X \in B_n\}$. We also know that

$$\bigcup_{n=1}^{\infty} E_n \equiv \bigcup_{n=1}^{\infty} \{X \in B_n\} = \left\{ X \in \bigcup_{n=1}^{\infty} B_n \right\},$$

where $\bigcup_{n=1}^{\infty} B_n \in \mathcal{B}(\mathbb{R})$, since $B_n \in \mathcal{B}(\mathbb{R})$ for every $n \geq 1$. This implies that also $\bigcup_{n=1}^{\infty} E_n \in \sigma(X)$ and completes the proof. □

Definition 413 *We call the σ-algebra defined by Equation (5.6) the* σ-algebra generated by *X.*

Theorem 414 *A function $X : \Omega \to \mathbb{R}$ is a real $\mathcal{E}$-random variable if and only if we have*

$$\sigma(X) \subseteq \mathcal{E}. \tag{5.7}$$

Proof. Also in this case, replacing $\mathcal{M}$ with $\mathcal{B}(\mathbb{R})$ the proof replies exactly the proof of Theorem 1420. □

Example 415 (Bernoulli random variable) *Fixed any pair of distinct points $x_0, x_1 \in \mathbb{R}$, say $x_0 < x_1$, assume there exist $E_0 \in \mathcal{E}$ such that*

$$0 < \mathbf{P}(E_0) < 1. \tag{5.8}$$

Equation (5.8) clearly imply

$$\varnothing \subset E_0 \subset \Omega. \tag{5.9}$$

Write $E_1 \equiv E_0^c$. Then, from Equation (5.9), it trivially follows

$$\varnothing \subset E_1 \subset \Omega. \tag{5.10}$$

As a consequence, the function $X : \Omega \to \mathbb{R}$ given by

$$X(\omega) \overset{def}{=} \begin{cases} x_0, & \text{if } \omega \in E_0, \\ x_1, & \text{if } \omega \in E_1, \end{cases} \tag{5.11}$$

is a real $\mathcal{E}$-random variable.

Discussion. Given any $B \in \mathcal{B}(\mathbb{R})$, only four cases are possible:

$$x_0 \in B \text{ and } x_1 \in B, \qquad x_0 \in B \text{ and } x_1 \notin B, \qquad x_0 \notin B \text{ and } x_1 \in B, \qquad x_0 \notin B \text{ and } x_1 \notin B.$$

Depending on which case occurs, we have

$$\{X \in B\} = \Omega, \quad \{X \in B\} = E_0, \quad \{X \in B\} = E_1, \quad \{X \in B\} = \varnothing.$$

Hence, in any case we have $\{X \in B\} \in \mathcal{E}$, which yields the desired result. □

Note that we have

$$\mathbf{P}(E_0) = \mathbf{P}(X = x_0), \qquad \mathbf{P}(E_1) = \mathbf{P}(X = x_1).$$

Definition 416 (Bernoulli real random variable) *We call the random variable defined by Equation (5.11) the* Bernoulli *real random variable with states x_0, x_1 and success probability $p \equiv \mathbf{P}(X = x_1)$. In case $x_0 \equiv 0$ and $x_1 \equiv 1$, we speak of the* standard Bernoulli *random variable and denote it by $Ber(p)$. In case $x_0 \equiv -1$ and $x_1 \equiv 1$, we speak of the* Rademacher *random variable and denote it by $Rad(p)$.*

Remark 417 (Bernoulli real random variable) *A function $X : \Omega \to \mathbb{R}$ is a Bernoulli random variable if and only if we have*

$$X = x_0 1_{E_0} + x_1 1_{E_0^c}$$

for a pair of distinct points $x_0, x_1 \in \mathbb{R}$, say $x_0 < x_1$, and an event $E_0 \in \mathcal{E}$ such that $0 < \mathbf{P}(E_0) < 1$.

Example 418 (indicator random variable) *Let E be event of Ω. Then, the indicator function $1_E : \Omega \to \mathbb{R}$ given by*

$$1_E(\omega) \overset{def}{=} \begin{cases} 1, & \text{if } \omega \in E, \\ 0, & \text{if } \omega \in E^c, \end{cases} \tag{5.12}$$

is a real $\mathcal{E}$-random variable if and only if $E \in \mathcal{E}$.

Discussion. We have

$$\{1\} \in \mathcal{B}(\mathbb{R}) \quad \text{and} \quad \{X = 1\} = E.$$

Hence, if the indicator function is a random variable, Then, $E \in \mathcal{E}$. Conversely, if $E \in \mathcal{E}$, Then, the argument is as in Example 415. □

Remark 419 (indicator random variable) *Assume $E \in \mathcal{E}$ and $0 < \mathbf{P}(E_0) < 1$. Then, the indicator function $1_E : \Omega \to \mathbb{R}$ is just a standard Bernoulli random variable.*

Definition 420 *We say that a function $X : \Omega \to \mathbb{R}$ is* discrete *if X takes a countable number of values. In symbols,*

$$|X(\Omega)| \leq \aleph_0. \tag{5.13}$$

Equivalently,

$$X(\Omega) \equiv \{x_n\}_{n \in N}, \quad N \subseteq \mathbb{N}. \tag{5.14}$$

Remark 421 *If a function $X : \Omega \to \mathbb{R}$ is discrete, Then, we have*

$$X = \sum_{n \in N} x_n 1_{E_n}, \tag{5.15}$$

where $E_n \equiv \{X = x_n\}$, for every $n \in N$.

Proposition 422 *If a function $X : \Omega \to \mathbb{R}$ is discrete, Then, we have*

$$\sigma\left(\{E_n\}_{n \in N}\right) = \sigma(X), \tag{5.16}$$

where $E_n \equiv \{X = x_n\}$, for every $n \in N$.

Proof. Under the assumption that $X : \Omega \to \mathbb{R}$ is discrete, that is Equation (5.14) holds true, the family $\{E_n\}_{n \in N}$, where $E_n \equiv \{X = x_n\}$, for every $n \in N$, is a partition of Ω. Hence, Proposition 250 applies and an event $E \in \sigma\left(\{E_n\}_{n \in N}\right)$ if and only if

$$E = \bigcup_{n \in N} E_n \tag{5.17}$$

for a suitable $K \subseteq N$. On the other hand, we have

$$\bigcup_{n \in K} E_n = \bigcup_{n \in K} \{X = x_n\} = \left\{ X \in \bigcup_{n \in K} \{x_n\} \right\} \tag{5.18}$$

where $\bigcup_{n \in K} \{x_n\} \in \mathcal{B}(\mathbb{R})$. Combining (5.17) and (5.18), it follows that $E \in \mathcal{B}(\mathbb{R})$, which implies

$$\sigma\left(\{E_n\}_{n \in N}\right) \subseteq \sigma(X).$$

Conversely, assume that $E \in \sigma(X)$, Then, $E = \{X \in B\}$ for some $B \in \mathcal{B}(\mathbb{R})$. On the other hand, under the assumption that $X : \Omega \to \mathbb{R}$ is discrete, we can write

$$\begin{aligned}\{X \in B\} &= \{X \in B\} \cap \Omega = \{X \in B\} \cap \{X \in X(\Omega)\} = \{X \in (B \cap X(\Omega))\} \\ &= \left\{X \in \left(B \cap \bigcup_{n \in N} \{x_n\}\right)\right\} = \left\{X \in \left(\bigcup_{n \in N} B \cap \{x_n\}\right)\right\} = \left\{X \in \bigcup_{n \in K} \{x_n\}\right\} \\ &= \bigcup_{n \in K} \{X = x_n\}.\end{aligned}$$

where $K = \{n \in N : x_n \in B\}$. It follows that $\{X \in B\} \in \sigma\left(\{E_n\}_{n \in N}\right)$. This shows that

$$\sigma(X) \subseteq \sigma\left(\{E_n\}_{n \in N}\right)$$

and completes the proof. □

Corollary 423 *If a function $X : \Omega \to \mathbb{R}$ is discrete, Then, X is a real $\mathcal{E}$-random variable if and only if*

$$E_n \in \mathcal{E}, \quad \forall n \in N, \tag{5.19}$$

where $E_n \equiv \{X = x_n\}$, for every $n \in N$.

Proof. If Equation (5.19) holds true, the characterization of $\sigma\left(\{E_n\}_{n \in N}\right)$ implies that $\sigma\left(\{E_n\}_{n \in N}\right) \subseteq \mathcal{E}$. On the other hand, thanks to Proposition 422 we have

$$\sigma\left(\{E_n\}_{n \in N}\right) = \sigma(X).$$

It follows

$$\sigma(X) \subseteq \mathcal{E}.$$

By virtue of Theorem 414, we Then, obtain that X is a real $\mathcal{E}$-random variable. The converse immediately follows considering that $E_n \equiv \{X = x_n\} \in \mathcal{B}(\mathbb{R})$, for every $n \in N$. □

Applying Corollary 423 we can build several discrete real $\mathcal{E}$-random variables.

Let $X : \Omega \to \mathbb{R}$ be a discrete real $\mathcal{E}$-random variable and let $P_X : \mathcal{B}(\mathbb{R}) \to \mathbb{R}_+$ be the distribution of X.

Example 424 (discrete uniform real random variable) *Fixed any $n \in \mathbb{N}$, $n \geq 3$, and any $x_1, \ldots, x_n \in \mathbb{R}$, say, such that $x_1 < \cdots < x_n$, assume there exists a sequence $(E_k)_{k=1}^n$ in $\mathcal{E}$ which is a partition of Ω. Assume further that*

$$\mathbf{P}(E_k) \equiv \frac{1}{n}, \quad \forall k = 1, \ldots, n, \tag{5.20}$$

Then, the function $X : \Omega \to \mathbb{R}$ given by

$$X(\omega) \overset{def}{=} x_k, \quad \forall \omega \in E_k, \ k = 1, \ldots, n, \quad \textit{equivalently} \quad X \overset{def}{=} \sum_{k=1}^{n} x_k 1_{E_k}, \tag{5.21}$$

is a real $\mathcal{E}$-random variable.

Definition 425 (discrete uniform real random variable) *We call the random variable defined by Equations (5.20) and (5.21) the* discrete uniform *real random variable with states* $x_1, \ldots, x_n$ *and denote it by* $Unif(x_1, \ldots, x_n)$. *In case* $x_k \equiv k$, *for* $k = 1, \ldots, n$, *we speak of the* standard discrete uniform *random variable and denote it by* $Unif(n)$.

Example 426 (binomial real random variable) *Fixed any* $n \in \mathbb{N}$, $n \geq 3$, *and any* $x_0, x_1, \ldots, x_n \in \mathbb{R}$, *say, such that* $x_0 < x_1 < \cdots < x_n$, *assume there exists a sequence* $(E_k)_{k=0}^n$ *in* $\mathcal{E}$ *which is a partition of* Ω. *Assume further that*

$$\mathbf{P}(E_k) \equiv \binom{n}{k} p^k q^{n-k}, \quad \forall k = 0, 1, \ldots, n, \tag{5.22}$$

where $p \in (0,1)$ *and* $q \equiv 1 - p$. *Then, the function* $X : \Omega \to \mathbb{R}$ *given by*

$$X(\omega) \stackrel{def}{=} x_k, \quad \forall \omega \in E_k,\ k = 0, 1, \ldots, n, \quad \textit{equivalently} \quad X \stackrel{def}{=} \sum_{k=0}^{n} x_k 1_{E_k}, \tag{5.23}$$

is a real $\mathcal{E}$*-random variable.*

Discussion. . □

Definition 427 (binomial real random variable) *We call the random variable defined by Equations (5.22) and (5.23) the* binomial *real random variable with states* $x_0, x_1, \ldots, x_n$ *and success probability* $p \equiv \mathbf{P}(X = x_1)$. *In case* $x_k \equiv k$, *for* $k = 0, 1 \ldots, n$, *we speak of the* standard binomial *random variable with* number of trials *[resp.* success probability*] parameter* n *[resp.* p*] and denote it by* $Bin(n, p)$.

Example 428 (continuous uniform real random variable) *Fixed any* $a, b \in \mathbb{R}$, *such that* $a < b$, *assume that*

$$\Omega \equiv [a, b], \quad \mathcal{E} \equiv \mathcal{B}([a, b]), \quad \mathbf{P} \equiv \frac{1}{b - a} \mu_L, \tag{5.24}$$

where $\mu_L : \mathcal{B}([a, b]) \to \mathbb{R}_+$ *is the Lebesgue measure on* $[a, b]$. *Then, the function* $X : \Omega \to \mathbb{R}$ *given by*

$$X(\omega) \stackrel{def}{=} \omega, \quad \forall \omega \in \Omega \tag{5.25}$$

is real $\mathcal{E}$*-random variable.*

Definition 429 (continuous uniform real random variable) *We call the random variable defined by Equations (5.24) and (5.24) the* continuous uniform *real random variable with states in* $[a, b]$ *and denote it by* $Unif(a, b)$. *In case* $a \equiv 0$ *and* $b \equiv 1$, *we speak of the* standard continuous uniform *random variable and denote it by* $Unif(0, 1)$.

Proposition 430 *Assume* $\mathcal{B}$ *is a basis for* $\mathcal{B}(\mathbb{R})$, *that is* $\sigma(\mathcal{B}) = \mathcal{B}(\mathbb{R})$. *Then, a function* $X : \Omega \to \mathbb{R}$ *is a real* $\mathcal{E}$-random variable *if and only the* X*-inverse image of any measurable set in* $\mathcal{B}$ *is an event in* $\mathcal{E}$. *In symbols*

$$\{X \in B\} \in \mathcal{E}, \quad \forall B \in \mathcal{B}. \tag{5.26}$$

Corollary 431 *A function $X : \Omega \to \mathbb{R}$ is a real $\mathcal{E}$-random variable if and only if*

$$\{X \in I\} \in \mathcal{E}, \quad \forall I \in \mathcal{I}, \tag{5.27}$$

were $\mathcal{I}$ is any of the families $\mathcal{I}_{o.\mathbb{Q}}(\mathbb{R})$, $\mathcal{I}_{c.\mathbb{Q}}(\mathbb{R})$, $\mathcal{I}_{o.c.\mathbb{Q}}(\mathbb{R})$, $\mathcal{I}_{c.o.\mathbb{Q}}(\mathbb{R})$ introduced in Proposition **??**.

Notation 432 *We will use the standard notations $\{a < X < b\}$, $\{a \leq X \leq b\}$, $\{a < X \leq b\}$, and $\{a \leq X < b\}$ as abbreviations for the events $\{X \in (a,b)\}$, $\{X \in [a,b]\}$, $\{X \in [a,b)\}$, and $\{X \in (a,b]\}$, respectively, for all $a, b \in \mathbb{R}$ such that $a < b$.*

Corollary 433 *The function $X : \Omega \to \mathbb{R}$ is a real $\mathcal{E}$-random variable if and only if*

$$\{X \in H\} \in \mathcal{E}, \quad \forall H \in \mathcal{H}, \tag{5.28}$$

were $\mathcal{H}$ is any of the families $\mathcal{H}_{r.o.\mathbb{Q}}(\mathbb{R})$, $\mathcal{H}_{r.c.\mathbb{Q}}(\mathbb{R})$, $\mathcal{H}_{l.o.\mathbb{Q}}(\mathbb{R})$, $\mathcal{H}_{l.c.\mathbb{Q}}(\mathbb{R})$ introduced in Proposition **??**.

Notation 434 *We will use the standard notations $\{X < a\}$, $\{X \leq a\}$, $\{X > a\}$, and $\{X \geq a\}$ as abbreviations for the events $\{X \in (-\infty, a)\}$, $\{X \in (-\infty, a]\}$, $\{X \in (a, +\infty)\}$, and $\{X \in [a, +\infty)\}$, respectively, for every $a \in \mathbb{R}$.*

Proposition 435 *Let $X : \Omega \to \mathbb{R}$ be a real $\mathcal{E}$-random variable and let $g : \mathbb{R} \to \mathbb{R}$ be a Borel function. Then, the function $g \circ X : \Omega \to \mathbb{R}$ given by*

$$(g \circ X)(\omega) \stackrel{def}{=} g(X(\omega)), \quad \forall \omega \in \Omega \tag{5.29}$$

is a random variable (see Proposition 1422). More generally, let $X_1 : \Omega \to \mathbb{R}, \ldots, X_N : \Omega \to \mathbb{R}$ be real $\mathcal{E}$-random variables, for some $N \in \mathbb{N}$, and let $h : \mathbb{R}^N \to \mathbb{R}$ be a Borel function. Then, the function $h \circ (X_1, \ldots, X_N) : \Omega \to \mathbb{R}$ given by

$$(h \circ (X_1, \ldots, X_N))(\omega) \stackrel{def}{=} h((X_1(\omega), \ldots, X_N(\omega))), \quad \forall \omega \in \Omega \tag{5.30}$$

is a real $\mathcal{E}$-random variable (see Proposition 1423).

Proof. First, observe that the subfamily of $\mathcal{P}(\mathbb{R}^N)$

$$\widetilde{\mathcal{B}} \equiv \{A \in \mathcal{P}(\mathbb{R}^N) \mid \{(X_1, \ldots, X_N) \in A\} \in \mathcal{E}\}$$

is a σ-algebra of $\mathbb{R}^N$. Actually, $\widetilde{\mathcal{B}}$ is nonempty, since $\mathbb{R}^N \in \widetilde{\mathcal{B}}$, for

$$\{(X_1, \ldots, X_N) \in \mathbb{R}^N\} = \Omega.$$

Furthermore, if $A \in \widetilde{\mathcal{B}}$, Then, also $A^c \in \widetilde{\mathcal{B}}$. In fact,

$$\{(X_1, \ldots, X_N) \in A^c\} = \{(X_1, \ldots, X_N) \in A\}^c.$$

In the end, if the sequence $(A_n)_{n \geq 1}$ belongs to $\widetilde{\mathcal{B}}$ Then, also $\bigcup_{n \geq 1} A_n$ belongs to $\widetilde{\mathcal{B}}$. In fact,

$$\left\{(X_1, \ldots, X_N) \in \bigcup_{n \geq 1} A_n\right\} = \bigcup_{n \geq 1} \{(X_1, \ldots, X_N) \in A_n\}.$$

Second, observe that for any Borel subset of $\mathbb{R}^N$ of the form $\mathsf{X}_{k=1}^N B_k$, where B_k is any Borel subset of $\mathbb{R}$, for $k = 1, \ldots, N$, we have

$$\left\{(X_1, \ldots, X_N) \in \mathsf{X}_{k=1}^N B_k\right\} = \{X_1 \in B_1\} \cap \ldots \cap \{X_N \in B_N\} \in \mathcal{E}.$$

Therefore, $\widetilde{\mathcal{B}}$ contains all the subsets of $\mathbb{R}^N$ of the form $\mathsf{X}_{k=1}^N B_k$. Combining the first and second observation, it follows that $\widetilde{\mathcal{B}}$ contains $\mathcal{B}(\mathbb{R}^N)$. In symbols,

$$\{(X_1, \ldots, X_N) \in B\} \in \mathcal{E}, \quad \forall B \in \mathcal{B}(\mathbb{R}^N). \tag{5.31}$$

Finally, since for any $B \in \mathcal{B}(\mathbb{R})$ the set $h^{-1}(B) \in \mathcal{B}(\mathbb{R}^N)$, combining the equality

$$\{h(X_1, \ldots, X_N) \in B\} = \{(X_1, \ldots, X_N) \in h^{-1}(B)\}$$

with (5.31) the desired claim follows. □

Thanks to the above result it is possible to prove that a large number of real functions on Ω, and a large number of their combination are random variables. For instance, under the assumption that $X : \Omega \to \mathbb{R}$ is a real $\mathcal{E}$-random variable we obtain

Example 436 *For any $\alpha \in \mathbb{R}$, the* multiple *of X with real coefficient α, that is the function $\alpha X : \Omega \to \mathbb{R}$ given by*

$$(\alpha X)(\omega) \overset{def}{=} \alpha X(\omega), \quad \forall \omega \in \Omega,$$

is a real $\mathcal{E}$-random variable.

Example 437 *For any $n \in \mathbb{N}$, the* power *of X with positive integer exponent n, that is the function $X^n : \Omega \to \mathbb{R}$ given by*

$$X^n(\omega) \overset{def}{=} X(\omega)^n, \quad \forall \omega \in \Omega,$$

is a real $\mathcal{E}$-random variable.

Example 438 *For any $n \in \mathbb{N}$ and any $a_0, a_1, \ldots, a_n \in \mathbb{R}$, the* polinomial *of degree n in X with coefficients $a_0, a_1, \ldots, a_n$, that is the function $\sum_{k=0}^n a_k X^k : \Omega \to \mathbb{R}$ given by*

$$\left(\sum_{k=0}^n a_k X^k\right)(\omega) \overset{def}{=} \sum_{k=0}^n a_k X^k(\omega) \equiv a_0 + a_1 X(\omega) + \cdots + a_{n-1} X^{n-1}(\omega) + a_n X^n(\omega), \quad \forall \omega \in \Omega,$$

is a real $\mathcal{E}$-random variable.

Example 439 *The* absolute value *of X, that is the function $|X| : \Omega \to \mathbb{R}$ given by*

$$|X|(\omega) \overset{def}{=} |X(\omega)|, \quad \forall \omega \in \Omega,$$

is a real $\mathcal{E}$-random variable.

Example 440 *The* positive *[resp.* negative*] part of X, that is the function $X^+ : \Omega \to \mathbb{R}$ [resp. $X^- : \Omega \to \mathbb{R}$] given by*

$$X^+(\omega) \overset{def}{=} \frac{1}{2}(|X(\omega)| + X(\omega)) \quad \text{[resp. } X^-(\omega) \overset{def}{=} \frac{1}{2}(|X(\omega)| - X(\omega))\text{]}, \quad \forall \omega \in \Omega,$$

is a real $\mathcal{E}$-random variable.

Example 441 *The* exponential *of X, that is the function $\exp \circ X : \Omega \to \mathbb{R}$ given by*

$$(\exp \circ X)(\omega) \overset{def}{=} \exp(X(\omega)), \quad \forall \omega \in \Omega,$$

is a real $\mathcal{E}$-random variable.

Example 442 *If $\mathbf{P}(X(\omega) \leq 0) = 0$, the* natural logarithm *of X, that is the function $\ln \circ X : \Omega \to \mathbb{R}$ given by*

$$(\ln \circ X)(\omega) \overset{def}{=} \begin{cases} \ln(X(\omega)), & \text{if } \omega \in \Omega - \{X(\omega) \leq 0\}, \\ \text{arbitrary}, & \text{if } \omega \in \{X(\omega) \leq 0\}, \end{cases}$$

is a real $\mathcal{E}$-random variable.

Example 443 *If $\mathbf{P}(X(\omega) \leq 0) = 0$, the* power *of X with real exponent α, that is the function $X^{\alpha} : \Omega \to \mathbb{R}$ given by*

$$X^{\alpha}(\omega) \overset{def}{=} \begin{cases} \exp(\alpha \ln(X(\omega))), & \text{if } \omega \in \Omega - \{X(\omega) \leq 0\}, \\ \text{arbitrary}, & \text{if } \omega \in \{X(\omega) \leq 0\}, \end{cases}$$

is a real $\mathcal{E}$-random variable.

Example 444 *The* sinus *[resp.* cosinus*] of X, that is the function $\sin X : \Omega \to \mathbb{R}$ [resp. $\cos X : \Omega \to \mathbb{R}$] given by*

$$\sin X(\omega) \overset{def}{=} \sin(X(\omega)) \qquad [\text{resp. } \cos X(\omega) \overset{def}{=} \cos(X(\omega))],$$

is a real $\mathcal{E}$-random variable.

Let X, Y be real $\mathcal{E}$-random variables we have

Example 445 *For all $a, b \in \mathbb{R}$ the* linear combination *of X and Y with coefficients a and b, that is the function $aX + bY : \Omega \to \mathbb{R}$, given by*

$$(aX + bY)(\omega) \overset{def}{=} aX(\omega) + bY(\omega), \quad \forall \omega \in \Omega,$$

is a real $\mathcal{E}$-random variable.

Example 446 *The* product *of X and Y, that is the function $XY : \Omega \to \mathbb{R}$, given by*

$$(XY)(\omega) \overset{def}{=} X(\omega) Y(\omega), \quad \forall \omega \in \Omega,$$

is a real $\mathcal{E}$-random variable.

Example 447 *If $\mathbf{P}(Y(\omega) = 0) = 0$, the* quotient *of X and Y, that is the function $X/Y : \Omega \to \mathbb{R}$ given by*

$$(X/Y)(\omega) \overset{def}{=} X(\omega) / Y(\omega), \quad \forall \omega \in \Omega,$$

is a real $\mathcal{E}$-random variable.

Example 448 *The* maximum *[resp.* minimum*] of X and Y, that is the function $X \vee Y : \Omega \to \mathbb{R}$ [resp. $X \wedge Y : \Omega \to \mathbb{R}$][1] given by*

$$(X \vee Y)(\omega) \overset{def}{=} \max\{X(\omega), Y(\omega)\} \quad [resp.\ (X \wedge Y)(\omega) \overset{def}{=} \min\{X(\omega), Y(\omega)\}], \quad \forall \omega \in \Omega,$$

is a real $\mathcal{E}$-random variable.

Discussion. Note that, we have

$$\{X \vee Y \leq z\} = \{X \leq z\} \cap \{Y \leq z\}, \quad \forall z \in \mathbb{R},$$

and

$$\{X \wedge Y \leq z\} = \{X \leq z\} \cup \{Y \leq z\}, \quad \forall z \in \mathbb{R}.$$

Similarly,

$$\{X \vee Y > z\} = \{X > z\} \cup \{Y > z\}, \quad \forall z \in \mathbb{R},$$

and

$$\{X \wedge Y > z\} = \{X > z\} \cap \{Y > z\}, \quad \forall z \in \mathbb{R}.$$

In fact,

$$\omega \in \{X \vee Y \leq z\} \Leftrightarrow X(\omega) \leq z \wedge Y(\omega) \leq z \Leftrightarrow \omega \in \{X \leq z\} \cap \{Y \leq z\},$$

and

$$\omega \in \{X \wedge Y \leq z\} \Leftrightarrow X(\omega) \leq z \vee Y(\omega) \leq z \Leftrightarrow \omega \in \{X \leq z\} \cup \{Y \leq z\}.$$

Similarly,

$$\omega \in \{X \vee Y > z\} \Leftrightarrow X(\omega) > z \vee Y(\omega) > z \Leftrightarrow \omega \in \{X > z\} \cup \{Y > z\},$$

and

$$\omega \in \{X \wedge Y > z\} \Leftrightarrow X(\omega) > z \wedge Y(\omega) > z \Leftrightarrow \omega \in \{X > z\} \cap \{Y > z\}.$$

Therefore, applying Corollary 433, the desired claim immediately follows. □

For any $n \in \mathbb{N}$, let $(X_k)_{k=1}^n$ be a sequence of n real $\mathcal{E}$-random variables. We have

Example 449 *For all $a_1, \ldots, a_n \in \mathbb{R}$ the* linear combination *of $X_1, \ldots, X_n$ with coefficients $a_1, \ldots, a_n$, that is the function $\sum_{k=1}^n a_k X_k : \Omega \to \mathbb{R}$, given by*

$$\left(\sum_{k=1}^n a_k X_k\right)(\omega) \overset{def}{=} \sum_{k=1}^n a_k X_k(\omega), \quad \forall \omega \in \Omega,$$

is a real $\mathcal{E}$-random variable.

Example 450 *For all $a_{k_1,\ldots,k_n} \in \mathbb{R}$ on varying of $k_1, \ldots, k_n \in \mathbb{N}_0$ such that $0 \leq k_1 + \cdots + k_n \leq n$, the* polinomial *of degree n in $X_1, \ldots, X_n$ with coefficients $a_{k_1,\ldots,k_n}$, that is the function $\sum_{k=0}^n \sum_{k_1,\ldots,k_n \in \mathbb{N}_0,\ k_1+\cdots+k_n=k} a_{k_1,\ldots,k_n} X_1^{k_1} + \cdots + X_n^{k_n} : \Omega \to \mathbb{R}$ given by*

$$\left(\sum_{k=0}^n \sum_{\substack{k_1,\ldots,k_n \in \mathbb{N}_0,\\ k_1+\cdots+k_n=k}} a_{k_1,\ldots,k_n} X_1^{k_1} + \cdots + X_n^{k_n}\right)(\omega) \overset{def}{=} \sum_{k=0}^n \sum_{\substack{k_1,\ldots,k_n \in \mathbb{N}_0,\\ k_1+\cdots+k_n=k}} a_{k_1,\ldots,k_n} X_1^{k_1}(\omega) + \cdots + X_n^{k_n}(\omega), \quad \forall \omega \in \Omega,$$

is a real $\mathcal{E}$-random variable.

[1] Another common notation for $X \vee Y$ [resp. $X \wedge Y$] is $\max(X, Y)$ [resp. $\min(X, Y)$].

Example 451 *The* maximum *of $X_1, \ldots, X_n$, that is the function $\bigvee_{k=1}^{n} X_k : \Omega \to \mathbb{R}^2$ given by*

$$\left(\bigvee_{k=1}^{n} X_k\right)(\omega) \overset{def}{=} \max\{X_1(\omega), \ldots, X_n(\omega)\}, \quad \forall \omega \in \Omega,$$

is a real $\mathcal{E}$-random variable.

Example 452 *The* minimum *of $X_1, \ldots, X_n$, that is the function $\bigwedge_{k=1}^{n} X_k : \Omega \to \mathbb{R}^3$ given by*

$$\left(\bigwedge_{k=1}^{n} X_k\right)(\omega) \overset{def}{=} \min\{X_1(\omega), \ldots, X_n(\omega)\}, \quad \forall \omega \in \Omega,$$

is a real $\mathcal{E}$-random variable.

5.2.1 Distribution of a Real Random Variable

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space and let $X : \Omega \to \mathbb{R}$, briefly X, be a real random variable on Ω. The distribution of X is just the distribution in the sense of Definition 1425 for $(\mathbb{X}, \mathcal{M}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. However, for the reader's convenience we restate the definition.

Notation 453 *For any $B \in \mathcal{B}(\mathbb{R})$, the notation $\mathbf{P}(X \in B)$ is a standard abbreviation for $\mathbf{P}(\{X \in B\})$. In particular, for any $x \in \mathbb{R}$, the notation $\mathbf{P}(X = x)$ is an abbreviation for $\mathbf{P}(\{X = x\})$, and for any $x, y \in \mathbb{R}$ such that $x \leq y$ the notations $\mathbf{P}(x < X \leq y)$, $\mathbf{P}(x < X < y)$, $\mathbf{P}(x \leq X \leq y)$, and $\mathbf{P}(x \leq X < y)$ are standard abbreviations for $\mathbf{P}(\{x < X \leq y\})$, $\mathbf{P}(\{x < X < y\})$, $\mathbf{P}(\{x \leq X \leq y\})$, and $\mathbf{P}(\{x \leq X < y\})$, respectively.*

Proposition 454 *Given any $\mathcal{E}$-random variable, $X : \Omega \to \mathbb{R}$, the map $P_X : \mathcal{B}(\mathbb{R}) \to \mathbb{R}_+$ given by*

$$P_X(B) \overset{def}{=} \mathbf{P}(X \in B), \quad \forall B \in \mathcal{B}(\mathbb{R}), \tag{5.32}$$

where $\mathbf{P}(X \in B)$ is the standard shorthand for $\mathbf{P}(\{X \in B\})$, is a probability distribution on $\mathbb{R}$.

Proof. We clearly have

$$\{X \in \mathbb{R}\} \equiv \{\omega \in \Omega : X(\omega) \in \mathbb{R}\} = \Omega.$$

In addition, given any sequence $(B_n)_{n \geq 1}$ of pairwise disjoint sets in $\mathcal{B}(\mathbb{R})$, the sequence $(\{X \in B_n\})_{n \geq 1}$ is a sequence of pairwise incompatible events in $\mathcal{E}$ such that

$$\left\{X \in \bigcup_{n=1}^{\infty} B_n\right\} = \bigcup_{n=1}^{\infty} \{X \in B_n\}.$$

Therefore, according Equation (21.9), we have

$$P_X(\mathbb{R}) = \mathbf{P}(X \in \mathbb{R}) = \mathbf{P}(\Omega) = 1.$$

Moreover,

$$P_X\left(\bigcup_{n=1}^{\infty} B_n\right) = \mathbf{P}\left(X \in \bigcup_{n=1}^{\infty} B_n\right) = \mathbf{P}\left(\bigcup_{n=1}^{\infty} \{X \in B_n\}\right) = \sum_{n=1}^{\infty} \mathbf{P}(X \in B_n) = \sum_{n=1}^{\infty} P_X(B_n).$$

Hence, Properties 1 and 2 are satisfied. This proves the claim. □

[2] Another common notation for $\bigvee_{k=1}^{n} X_k$ is $\max(X_1, \ldots, X_n)$.
[3] Another common notation for $\bigwedge_{k=1}^{n} X_k$ is $\min(X_1, \ldots, X_n)$.

Definition 455 *We call the* distribution *of the real random variable X the probability distribution $P_X : \mathcal{B}(\mathbb{R}) \to \mathbb{R}_+$, briefly P_X, given by Equation 5.32.*

Example 456 *Given any $x_0 \in \mathbb{R}$, let X be the Dirac real random variable concentrated at x_0 (see Definition 406) and let $P_X : \mathcal{B}(\mathbb{R}) \to \mathbb{R}_+$ be the distribution of X. We have*

$$P_X(B) = \begin{cases} 1, & \text{if } x_0 \in B, \\ 0, & \text{if } x_0 \notin B, \end{cases} \qquad \forall B \in \mathcal{B}(\mathbb{R}).$$

The distribution of the Dirac real random variable concentrated at x_0 is the Dirac probability distribution on $\mathbb{R}$ concentrated at x_0 (see Definition 263).

Example 457 *Given any $x_0, x_1 \in \mathbb{R}$ such that $x_0 < x_1$, let X be the Bernoulli random variable with states x_0, x_1 and success probability $p \equiv \mathbf{P}(X = x_1)$ (see Definition 416). Let $P_X : \mathcal{B}(\mathbb{R}) \to \mathbb{R}_+$ be the distribution of X. We have*

$$P_X(B) = \begin{cases} 1, & \text{if } x_0 \in B \text{ and } x_1 \in B, \\ q, & \text{if } x_0 \in B \text{ and } x_1 \notin B, \\ p, & \text{if } x_0 \notin B \text{ and } x_1 \in B, \\ 0, & \text{if } x_0 \notin B \text{ and } x_1 \notin B. \end{cases}$$

The distribution of the Bernoulli random variable with states x_0, x_1 and success probability p is the Bernoulli probability distribution on $\mathbb{R}$ with success probability p (see Definition 266).

Example 458 *Given $n \in \mathbb{N}$, $n \geq 3$, and given $x_0, x_1, \ldots, x_n \in \mathbb{R}$ such that $x_0 < x_1 < \cdots < x_n$, let X be the binomial random variable with states $x_0, x_1, \ldots, x_n$ and success probability $p \equiv \mathbf{P}(X = x_1)$ (see Definition 427). Let $P_X : \mathcal{B}(\mathbb{R}) \to \mathbb{R}_+$ be the distribution of X. We have*

$$P_X(B) = \sum_{k \in \{0,1,\ldots,n \,:\, x_k \in B\}} \binom{n}{k} p^k q^{n-k}.$$

The distribution of the binomial random variable with states $x_0, x_1, \ldots, x_n$ and success probability p is the binomial probability distribution on $\mathbb{R}$ with success probability p (see Definition 272).

Symmetric Random Variables

Let $X : \Omega \to \mathbb{R}$, briefly X, be a real random variable on a probability space Ω.

Definition 459 *We say that X is symmetric about a state $x_0 \in \mathbb{R}$ if*

$$P_{X - x_0} = P_{x_0 - X}, \tag{5.33}$$

where P_{X-x_0} and P_{x_0-X} are the probability distribution of the real random variables $X - x_0$ and $x_0 - X$, respectively.

Example 460 *Given any $x_0 \in \mathbb{R}$, let X be the Dirac real random variable concentrated at x_0 (see Definition 406). Then, X is symmetric about x_0.*

Discussion. We have

$$(X - x_0)(\omega) = X(\omega) - x_0 = 0 = x_0 - X(\omega) = (x_0 - X)(\omega).$$

Therefore, both $X - x_0$ and $x_0 - X$ turn out to be Dirac real random variables concentrated at 0. This clearly implies that (5.33) holds true and the claim follows. □

Example 461 *Given any $x_0, x_1 \in \mathbb{R}$ such that $x_0 < x_1$, let X be the Bernoulli random variable with states x_0, x_1 and success probability $p \equiv \mathbf{P}(X = x_1)$ (see Definition 416). Assume $p = 1/2$. Then, X is symmetric about $x_s \equiv (x_0 + x_1)/2$.*

Discussion. We have

$$(X - x_s)(\omega) = X(\omega) - x_s = X(\omega) - \frac{x_0 + x_1}{2} = \begin{cases} \frac{x_0 - x_1}{2} & \text{if } X(\omega) = x_0 \\ \frac{x_1 - x_0}{2} & \text{if } X(\omega) = x_1 \end{cases}$$

and

$$(x_s - X)(\omega) = x_s - X(\omega) = \frac{x_0 + x_1}{2} - X(\omega) = \begin{cases} \frac{x_1 - x_0}{2} & \text{if } X(\omega) = x_0 \\ \frac{x_0 - x_1}{2} & \text{if } X(\omega) = x_1 \end{cases}.$$

On the other hand, the assumption $p = 1/2$ implies

$$\mathbf{P}(X = x_0) = \mathbf{P}(X = x_1) = \frac{1}{2}.$$

As a consequence,

$$\mathbf{P}\left(X - x_s = \frac{x_0 - x_1}{2}\right) = \mathbf{P}\left(X - x_s = \frac{x_1 - x_0}{2}\right) = \frac{1}{2}$$

and

$$\mathbf{P}\left(x_s - X = \frac{x_1 - x_0}{2}\right) = \mathbf{P}\left(x_s - X = \frac{x_0 - x_1}{2}\right) = \frac{1}{2}.$$

Therefore, the random variables $X - x_s$ and $x_s - X$ take the same states $\frac{x_0 - x_1}{2}$ and $\frac{x_1 - x_0}{2}$ with the same probabilities. It easily follows that (5.33) holds true and we obtain the claim. □

Example 462 *Given any $x_0, x_1 \in \mathbb{R}$ such that $x_0 < x_1$, let X be the Bernoulli random variable with states x_0, x_1 and success probability $p \equiv \mathbf{P}(X = x_1)$ (see Definition 416). Assume $p \neq 1/2$. Then, X is not symmetric about any state $x_s \in \mathbb{R}$.*

Discussion. Assume there exist a state $x_s \in \mathbb{R}$ about which X is symmetric.Then,

$$P_{X - x_s}(B) = P_{x_s - X}(B)$$

for every $B \in \mathcal{B}(\mathbb{R})$. In particular,

$$P_{X - x_s}(x) = P_{x_s - X}(x) \tag{5.34}$$

for every $x \in \mathbb{R}$. On the other hand,

$$P_{X - x_s}(x) = \mathbf{P}(X - x_s = x) = \mathbf{P}(X = x_s + x)$$

and

$$P_{x_s - X}(x) = \mathbf{P}(x_s - X = x) = \mathbf{P}(X = x_s - x).$$

Now,

$$\mathbf{P}(X = x_s + x) = \begin{cases} q & \text{if } x_s + x = x_0 \Leftrightarrow x = x_0 - x_s \\ p & \text{if } x_s + x = x_1 \Leftrightarrow x = x_1 - x_s \\ 0 & \text{otherwise} \end{cases}.$$

Similarly,

$$\mathbf{P}(X = x_s - x) = \begin{cases} q & \text{if } x_s - x = x_0 \Leftrightarrow x = x_s - x_0 \\ p & \text{if } x_s - x = x_1 \Leftrightarrow x = x_s - x_1 \\ 0 & \text{otherwise} \end{cases}.$$

Hence, Equation (5.34) implies

$$x_0 - x_s = x_s - x_0 \quad \text{and} \quad x_1 - x_s = x_s - x_1. \tag{5.35}$$

From (5.35) it Then, follows

$$x_s = x_0 \quad \text{and} \quad x_s = x_1,$$

which is clerly impossible. We can Then, conlude that it cannot exist a state x_s about which X is symmetric. □

Proposition 463 *A real random variable X is symmetric about a state $x_0 \in \mathbb{R}$ if and only if we have*

$$\mathbf{P}(X \leq x_0 + x) = \mathbf{P}(X \geq x_0 - x), \tag{5.36}$$

for every $x \in \mathbb{R}$.

Proof. . □

5.2.2 Distribution Function of a Real Random Variable

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space, let $(\mathbb{R}, \mathcal{B}(\mathbb{R})) \equiv \mathbb{R}$ be the Euclidean real line equipped with the Borel σ-algebra, let $X : \Omega \to \mathbb{R}$ be a real random variable, and let $P_X : \mathcal{B}(\mathbb{R}) \to \mathbb{R}_+$ be the distribution of X (see Definition 1425).

Definition 464 *We call the* distribution function *of X the function $F_X : \mathbb{R} \to \mathbb{R}$ given by*

$$F_X(x) \overset{def}{=} \mathbf{P}(X \leq x), \quad \forall x \in \mathbb{R}. \tag{5.37}$$

Remark 465 *We clearly have*

$$F_X(x) = P_X((-\infty, x]), \tag{5.38}$$

for every $x \in \mathbb{R}$.

Example 466 *Fixed any $x_0 \in \mathbb{R}$, let X be the Dirac real random variable concentrated at x_0 (see Definition 406). Write $F_X : \mathbb{R} \to \mathbb{R}$ for the distribution function of X. We have*

$$F_X(x) = \begin{cases} 0, & \text{if } x < x_0, \\ 1, & \text{if } x_0 \leq x. \end{cases} \tag{5.39}$$

Equivalently,

$$F_X(x) = 1_{[x_0, +\infty)}(x). \tag{5.40}$$

Definition 467 *In case $x_0 = 0$, the distribution function given by (5.39) is referred to as the* Heavside step function*.*

Example 468 *Fixed any $x_0, x_1 \in \mathbb{R}$, say, such that $x_0 < x_1$, let $X : \Omega \to \mathbb{R}$ be the Bernoulli random variable with states x_0, x_1 and success probability $p \equiv \mathbf{P}(X = x_1)$ (see Definition 416). Write $F_X : \mathbb{R} \to \mathbb{R}$ for the distribution function of X. We have*

$$F_X(x) = \begin{cases} 0, & \text{if } x < x_0, \\ q, & \text{if } x_0 \leq x < x_1, \\ 1, & \text{if } x_1 \leq x. \end{cases} \tag{5.41}$$

Equivalently,

$$F_X(x) = q1_{[x_0,+\infty)}(x) + p1_{[x_1,+\infty)}(x). \tag{5.42}$$

Example 469 *Fixed any $n \in \mathbb{N}$, $n \geq 3$, and $x_1, \ldots, x_n \in \mathbb{R}$, say, such that $x_1 < \cdots < x_n$, let $X : \Omega \to \mathbb{R}$ be the uniform discrete random variable with states $x_1, \ldots, x_n$ (see Definition 425). Write $F_X : \mathbb{R} \to \mathbb{R}$ for the distribution function of X. We have*

$$F_X(x) = \begin{cases} 0 & \text{if } x < x_1 \\ \frac{k}{n} & \text{if } x_k \leq x < x_{k+1}, \quad \forall k = 1, \ldots, n-1 \\ 1 & \text{if } x_n \leq x \end{cases}. \tag{5.43}$$

Equivalently,

$$F_X(x) = \sum_{k=1}^{n} \frac{1}{n} 1_{[x_k,+\infty)}(x). \tag{5.44}$$

Example 470 *Given $n \in \mathbb{N}$, $n \geq 3$, and given $x_0, x_1, \ldots, x_n \in \mathbb{R}$ such that $x_0 < x_1 < \cdots < x_n$, let $X : \Omega \to \mathbb{R}$ be the binomial random variable with states $x_0, x_1, \ldots, x_n$ and success probability $p \equiv \mathbf{P}(X = x_1)$ (see Definition 427). Write $F_X : \mathbb{R} \to \mathbb{R}$ for the distribution function of X. We have*

$$F_X(x) = \begin{cases} 0 & \text{if } x < x_0 \\ \sum\limits_{j=0}^{k} \binom{n}{j} p^j q^{n-j} & \text{if } x_k \leq x < x_{k+1}, \quad \forall k = 0, 1, \ldots, n-1 \\ 1 & \text{if } x_n \leq x \end{cases}. \tag{5.45}$$

Equivalently,

$$F_X(x) = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} 1_{[x_k,+\infty)}(x). \tag{5.46}$$

Theorem 471 *Assume the random variable X takes a finite number of values, that is $X(\Omega) \equiv \{x_k\}_{k=1}^n$, for some $n \in \mathbb{N}$ and $x_1, \ldots, x_n \in \mathbb{R}$. Write $p_k \equiv \mathbf{P}(X = x_k)$, for every $k = 1, \ldots, n$. Then, we have*

$$F_X(x) = \sum_{k=1}^{n} p_k 1_{[x_k,+\infty)}(x), \tag{5.47}$$

for every $x \in \mathbb{R}$.

Proof. . □

From Theorem 471 and the previous example it is clearly seen that a distribution function is not necessarily continuous. Yet a distribution function has significant properties that we describe in what follows.

Let $F_X : \mathbb{R} \to \mathbb{R}$ be the distribution function of X.

Proposition 472 *The function F_X satisfies the following properties:*

1. $F_X(x) \geq 0$, *for every* $x \in \mathbb{R}$;
2. $\lim_{x \to -\infty} F_X(x) = 0$;
3. $\lim_{x \to +\infty} F_X(x) = 1$;
4. $F_X(x) \leq F_X(y)$, *for all* $x, y \in \mathbb{R}$ *such that* $x \leq y$;
5. $\exists \lim_{x \to x_0^-} F_X(x) \equiv \ell \in \mathbb{R}$, *with* $\ell \leq F_X(x_0)$, *for every* $x_0 \in \mathbb{R}$;
6. $\exists \lim_{x \to x_0^+} F_X(x) = F_X(x_0)$, *for every* $x_0 \in \mathbb{R}$, *namely* F_X *is right-continuous.*

Proof. . □

Proposition 473 *For all* $x \in \mathbb{R}$, *we have*

$$\lim_{u \to x^-} F_X(u) = \mathbf{P}(X < x). \tag{5.48}$$

Proof. Considering any increasing sequence $(x_n)_{n \geq 1}$ which converges to x from the left, we can write

$$\bigcup_{n \geq 1} (-\infty, x_n] = (-\infty, x).$$

Hence,

$$P_X\left(\bigcup_{n \geq 1} (-\infty, x_n]\right) = P_X(-\infty, x). \tag{5.49}$$

On the other hand, by virtue of Equation (4.43), we have

$$P_X\left(\bigcup_{n \geq 1} (-\infty, x_n]\right) = \lim_{n \to \infty} P_X((-\infty, x_n]) = \lim_{n \to \infty} F_X(x_n) \tag{5.50}$$

and since $(x_n)_{n \geq 1}$ converges to x from the left, considering 5 in Proposition 472, we obtain

$$\lim_{n \to \infty} F_X(x_n) = \lim_{u \to x^-} F_X(u). \tag{5.51}$$

Combining (5.49)-(5.51), it follows

$$P_X(-\infty, x) = \lim_{u \to x^-} F_X(u).$$

In the end, considering (5.32), we obtain the desired (5.48). □

Proposition 474 *For all* $x, y \in \mathbb{R}$ *such that* $x < y$, *we have*

$$\mathbf{P}(x < X \leq y) = F_X(y) - F_X(x). \tag{5.52}$$

Proof. Considering Equations (5.32) and (5.38), we can write

$$\begin{aligned}\mathbf{P}(x < X \leq y) &= P_X((x,y]) = P_X((-\infty,y] - (-\infty,x]) \\ &= P_X((-\infty,y]) - P_X((-\infty,x]) = F_X(y) - F_X(x),\end{aligned}$$

as claimed. □

Proposition 475 *For all $x, y \in \mathbb{R}$ such that $x < y$ we have*

$$\mathbf{P}(x \leq X \leq y) = F_X(y) - \lim_{u \to x} F_X(u). \tag{5.53}$$

Proof. Similarly to the proof of Proposition 474, we can write

$$\begin{aligned}\mathbf{P}(x \leq X \leq y) &= P_X([x,y]) = P_X((-\infty,y] - (-\infty,x)) \\ &= P_X((-\infty,y]) - P_X((-\infty,x)) = F_X(y) - P_X((-\infty,x)).\end{aligned}$$

Therefore, on account of Equation (5.48), the desired (5.53) follows. □

Proposition 476 *For all $x, y \in \mathbb{R}$ such that $x < y$ we have*

$$\mathbf{P}(x < X < y) = \lim_{v \to y} F_X(v) - F_X(x)$$

Proof. Similarly to the proof of Proposition 474, we can write

$$\begin{aligned}\mathbf{P}(x < X < y) &= P_X((x,y)) = P_X((-\infty,y) - (-\infty,x]) \\ &= P_X((-\infty,y)) - P_X((-\infty,x]) = P_X((-\infty,y)) - F_X(x).\end{aligned}$$

Therefore, thanks to Equation (5.48), the desired (**??**) follows. □

Proposition 477 *For all $x, y \in \mathbb{R}$ such that $x < y$ we have*

$$\mathbf{P}(x \leq X < y) = \lim_{v \to y} F_X(v) - \lim_{u \to x} F_X(u). \tag{5.54}$$

Proof. Similarly to the proof of Proposition 474, we can write.

$$\mathbf{P}(x \leq X < y) = P_X([x,y)) = P_X((-\infty,y) - (-\infty,x)) = P_X((-\infty,y)) - P_X((-\infty,x))$$

Therefore, thanks to Equation (5.48) the desired (5.54) follows. □

Proposition 478 *For every $x \in \mathbb{R}$ we have*

$$\mathbf{P}(X = x) = F_X(x) - \lim_{u \to x^-} F_X(u). \tag{5.55}$$

Proof. Similarly to the proof of (5.48), considering any increasing sequence $(x_n)_{n \geq 1}$ which converges to x, we can write

$$\{x\} = \bigcap_{n \geq 1} (x_n, x].$$

Hence,

$$P_X(x) = P_X\left(\bigcap_{n \geq 1} (x_n, x]\right). \tag{5.56}$$

On the other hand, by virtue of (4.43), (5.52), and 4 in Proposition 472, we have

$$P_X\left(\bigcap_{n\geq 1}(x_n, x]\right) = \lim_{n\to\infty} P_X((x_n, x]) = \lim_{n\to\infty}(F_X(x) - F_X(x_n))$$
$$= F_X(x) - \lim_{n\to\infty} F_X(x_n) = F_X(x) - \lim_{u\to x^-} F_X(u). \tag{5.57}$$

Combining (5.56) and (5.57), it follows

$$P_X(x) = F_X(x) - \lim_{u\to x^-} F_X(u).$$

In the end, on account of (5.32), we obtain the desired (5.55). □

Corollary 479 *If F_X is continuous, Then, for all $x, y \in \mathbb{R}$ such that $x < y$ we have*

$$\mathbf{P}(x < X \leq y) = \mathbf{P}(x < X < y) = \mathbf{P}(x \leq X \leq y) = \mathbf{P}(x \leq X < y). \tag{5.58}$$

Proof. . □

Proposition 480 *F_X is continuous if and only if for every $x \in \mathbb{R}$ we have*

$$\mathbf{P}(X = x) = 0. \tag{5.59}$$

Proof. . □

Proposition 481 *The function F_X is differentiable almost everywhere on $\mathbb{R}$ and we have $F'_X(x) \geq 0$ at any point $x \in \mathbb{R}$ where F_X is differentiable.*

Proof. An important result of Calculus states that any monotonic function defined on an interval of $\mathbb{R}$ is almost everywhere differentiable on the interval. As F_X is non-decreasing on $\mathbb{R}$ and we have

$$F'_X(x) \stackrel{\text{def}}{=} \lim_{\Delta x\to 0} \frac{F(x+\Delta x) - F(x)}{\Delta x},$$

provided the limit is a real number, for the theorem on persistence of sign we obtain $F'_X(x) \geq 0$ wherever defined. □

Definition 482 *We call a* derivative *of F_X any function $f_X : \mathbb{R} \to \mathbb{R}$ such that*

$$f_X(x) \stackrel{def}{=} \begin{cases} F'_X(x) & \text{if } F \text{ is differentiable in } x \\ \text{arbitrary} & \text{if } F \text{ is not differentiable in } x \end{cases}.$$

Let $F : \mathbb{R} \to \mathbb{R}$ be a function.

Definition 483 *We say that $F : \mathbb{R} \to \mathbb{R}$ is a* distribution, *if $F : \mathbb{R} \to \mathbb{R}$ satisfies Properties 1-6 in Proposition 472.*

Theorem 484 (II Inversion Theorem) *Given any distribution $F : \mathbb{R} \to \mathbb{R}$, there always exist a probability space $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ and a random variable $X : \Omega \to \mathbb{R}$ such that the distribution function $F_X : \mathbb{R} \to \mathbb{R}$ of X is the given $F : \mathbb{R} \to \mathbb{R}$.*

Proof. ... □

P-P plot

A P-P plot, where P-P stands for *probability-probability* or *percent-percent*, is a graphical method to compare two probability distributions by plotting their distribution functions against each other.

Let X, Y be real random variables with distribution function $F_X : \mathbb{R} \to \mathbb{R}$ and $F_Y : \mathbb{R} \to \mathbb{R}$, respectively.

Definition 485 *We call a* P-P plot *of Y against X, the representation in the Cartesian plane $\mathbb{R}^2$ of the parametric curve $PP_{X,Y} : \mathbb{R} \to \mathbb{R}^2$ given by*

$$PP_{X,Y}(u) \stackrel{def}{=} (F_X(u), F_Y(u)), \quad \forall u \in \mathbb{R}. \tag{5.60}$$

The shape of a P-P plot provides us with several information about the two distributions.

Remark 486 *Assume $Y = X$. Then, the pattern of the corresponding PP -plot $PP_{X,Y}$ lies on the straigth line $y = x$. Conversely, assume that the pattern of the P-P plot $PP_{X,Y}$ lies on the straigth line $y = x$. Then, $Y = X$.*

Remark 487 *Assume the test data set y is drawn from the distribution of X. Then, the pattern of the corresponding P-P plot $PP_{X,y}$ is very close to the straight line $y = x$. Conversely, assume that the pattern of the P-P plot $PP_{X,y}$ is very close to the straigth line $y = x$. Then, the test data set y is likely drawn from the distribution of X.*

Remark 488 *Assume the distribution of Y is more concentrated [resp. dispersed] than the distribution of X. Then, the pattern of the corresponding P-P plot $PP_{X,Y}$ is steeper [resp. flatter] than the straigth line $y = x$. Conversely, assume that the pattern of the P-P plot $PP_{X,Y}$ is steeper [resp. flatter] than the straigth line $y = x$. Then, the distribution of Y is more concentrated [resp. dispersed] than the distribution of X.*

Remark 489 *Assume the test data set y is drawn from a distribution which is more concentrated [resp. dispersed] than the distribution of X. Then, the pattern of the corresponding P-P plot $PP_{X,y}$ is steeper [resp. flatter] than the straigth line $y = x$. Conversely, assume that the pattern of the P-P plot $PP_{X,y}$ is steeper [resp. flatter] than the straigth line $y = x$. Then, the test data set y is likely drawn from a distribution which is more concentrated [resp. dispersed] than the distribution of X.*

Remark 490 *Assume the distribution of Y is more concentrated on the left [resp. right] than the distribution of X. Then, the pattern of the corresponding P-P plot $PP_{X,Y}$ is arched downwards [resp. upwards]. Conversely, assume that the pattern of the P-P plot $PP_{X,Y}$ is arched downwards [resp. upwards]. Then, the distribution of Y is more concentrated on the left [resp. right] than the distribution of X.*

Remark 491 *Assume the test data set y is drawn from a distribution which is more concentrated on the left [resp. right] than the distribution of X. Then, the pattern of the corresponding P-P plot $PP_{X,y}$ is arched downwards [resp. upwards]. Conversely, assume that the pattern of the P-P plot $PP_{X,y}$ arched downwards [resp. upwards]. Then, the test data set y is likely drawn from a distribution which is more concentrated on the left [resp. right] than the distribution of X.*

Remark 492 *Assume the distribution of Y has lighter [resp. heavier] tails than the distribution of X. Then, the pattern of the corresponding P-P plot $PP_{X,Y}$ is "S" [resp. reverse "S"]-shaped like the logistic [resp. logit]. Conversely, assume that the pattern of the P-P plot $PP_{X,Y}$ is "S" [resp. reverse "S"]-shaped like the logistic [resp. logit]. Then, the distribution of Y has lighter [resp. heavier] tails than the distribution of X.*

Remark 493 *Assume the test data set y is drawn from a distribution which has lighter [resp. heavier] tails than the distribution of X. Then, the pattern of the corresponding P-P plot $PP_{X,y}$ is "S" [resp. reverse "S"]-shaped like the logistic [resp. logit]. Conversely, assume that the pattern of the P-P plot $PP_{X,y}$ is "S" [resp. reverse "S"]-shaped like the logistic [resp. logit]. Then, the test data set y is likely drawn from a distribution which has lighter [resp. heavier] tails than the distribution of X.Let X be a real random variable with distribution function $F_X : \mathbb{R} \to \mathbb{R}$ and let $(y_k)_{k=1}^N \equiv y$ a data set of length $N \geq 1$ with order statistic $\left(y_{(k)}\right)_{k=1}^N$ and empirical distribution function $F_y : \mathbb{R} \to \mathbb{R}$.*

Definition 494 *We call the distribution function F_X [resp. the data set y] the* reference *distribution [resp. the* test *data set].*

Definition 495 *We call a* P-P plot *of y against X, the representation in the Cartesian plane $\mathbb{R}^2$ of the parametric curve $PP_{X,y} : \{1, \ldots, N\} \to \mathbb{R}^2$ given by*

$$PP_{X,y}(k) \stackrel{def}{=} \left(F_X\left(y_{(k)}\right), F_y\left(y_{(k)}\right)\right), \quad \forall k \in \{1, \ldots, N\}. \tag{5.61}$$

To better understanding the main features of P-P plots, it may be interesting to draw a P-P plot of the distribution function of some random variables against the distribution function of other random variables.

Assume X and Y are two random variables whose distribution functions are implemented in Matlab as "name$X$" and "name$Y$", respectively. Then, the commands

$$u = u_{\min} : step : u_{\max}; \quad Fx = cdf\,('NameX', u, params)\,; \quad Fy = cdf\,('NameY', u, params)\,;$$

generate the vectors Fx and Fy of the values taken by the distribution functions of X and Y, respectively, on the states u of the interval $[u_{\min}, u_{\max}]$ varying by step of *step*. Hence the desired P-P- plot $PP_{X,Y}$ is drawn from the commands

$$figure; \quad plot(Fx, Fy, '?')\,; \qquad \text{or} \qquad figure; \quad scatter(Fx, Fy, '?')\,;$$

where *?* is any dot identification marker among o, +, *, and others (see https://it.mathworks.com/help/matla
Note that in the command *plot* a dot identification marker is necessary if we are interested in drawing also the LS-line by adding the commands

$$hold\ oN; \quad rliNe = lsliNe; \quad set(rliNe, 'Color', '?');$$

where *?* is any colour identification marker among r, g, b, and others (see https://it.mathworks.com/help/mat
Otherwise it might be omitted.

More interesting is to draw the P-P of a data set $(y_k)_{k=1}^N \equiv y$ against a random variable X whose distribution function is implemented in Matlab as "name$X$". To this goal, we need to build the empirical distribution function F_y of the data set y by the commands

$$[Fy, uy] = ecdf\,(y)\,;$$

which returns the order statistic $y_{()} = uy$ of y and the empirical distribution function Fy of y evaluated on the order statistic $y_{()}$. Hence, the command

$$Fx = cdf\left('NameX', uy, params\right);$$

generates the vector Fx of the values taken by the distribution functions of X on the states uy. In the end, the desired P-P- plot is drawn from the commands

$$figure; \quad plot\left(x_{()}, y_{()}, '?\right); \qquad \text{or} \qquad figure; \quad scatter\left(x_{()}, y_{()}, '?\right);$$

where *?* is any dot identification marker among o, +, *, and others (see https://it.mathworks.com/help/matla

5.2.3 Density Function of a Real Random Variable

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space, let $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu_L) \equiv \mathbb{R}$ be the Euclidean real line equipped with the Borel σ-algebra and the Lebesgue measure μ_L. Let $X : \Omega \to \mathbb{R}$ be a real random variable and let $P_X : \mathcal{B}(\mathbb{R}) \to \mathbb{R}_+$ and $F_X : \mathbb{R} \to \mathbb{R}$ be the distribution and the distribution function of X, respectively.

We observed in Subsection 5.2.2 that F_X is a non-decreasing function. Hence, F_X is differentiable almost everywhere on $\mathbb{R}$. However, in general, F_X is not continuous. This clearly implies that, in general, F_X is not everywhere differentiable on $\mathbb{R}$. However, even in cases in which F_X is continuous, F_X is not necessarily everywhere differentiable on $\mathbb{R}$. To establish a useful probabilistic relationship between distribution functions and their derivatives we need to consider *absolutely continuous* distribution functions.

Definition 496 *We say that F_X is* absolutely continuous *on $\mathbb{R}$, if for every $\varepsilon > 0$ there exists δ_ε such that for any finite sequence $([x_n, y_n])_{n=1}^N$ of mutually disjoint interval of $\mathbb{R}$ fulfilling $\sum_{n=1}^N |y_n - x_n| < \delta_\varepsilon$ we have $\sum_{n=1}^N |F_X(y_n) - F_X(x_n)| < \varepsilon$.*

Remark 497 *If F_X is absolutely continuous on $\mathbb{R}$, Then, F_X is uniformly continuous. In particular, F_X is continuous.*

Definition 498 *We say that F_X is* Lipschitz continuous *on $\mathbb{R}$, if there exists $L > 0$ such that we have*

$$|F_X(y) - F_X(x)| < L|y - x|, \quad \forall x, y \in \mathbb{R}.$$

Proposition 499 *If F_X is Lipschitz continuous on $\mathbb{R}$, Then, F_X is absolutely continuous.*

Proposition 500 *If F_X is differentiable everywhere on $\mathbb{R}$ and the derivative $f_X : \mathbb{R} \to \mathbb{R}$ is bounded, Then, F_X is absolutely continuous.*

Theorem 501 *The distribution function F_X is absolutely continuous on $\mathbb{R}$, if and only if*

1. *any derivative $f_X : \mathbb{R} \to \mathbb{R}$ of F_X, in the sense of Definition 482, is Lebesgue integrable on $\mathbb{R}$;*
2. *we have*

$$F_X(x) = \int_{(-\infty, x)} f_X(u)\, d\mu_L(u), \quad \forall x \in \mathbb{R}. \tag{5.62}$$

Corollary 502 *The distribution function F_X of X is absolutely continuous on $\mathbb{R}$ if and only if the distribution P_X of X is absolutely continuous with respect to the Lebesgue measure μ_L on $\mathbb{R}$.*

Definition 503 *We say that the real random variable X is* absolutely continuous *if the distribution function F_X of X is absolutely continuous or equivalently if the distribution P_X of X is absolutely continuous with respect to the Lebesgue measure μ_L on $\mathbb{R}$. We have clearly*

Proposition 504 *If there exists a Lebesgue integrable function $f : \mathbb{R} \to \mathbb{R}$ such that*

$$F_X(x) = \int_{(-\infty,x)} f(u)\, d\mu_L(u), \quad \forall x \in \mathbb{R}, \tag{5.63}$$

Then, F_X is absolutely continuous on $\mathbb{R}$ and we have

$$f_X(x) = f(x), \quad a.e. \text{ on } \mathbb{R},$$

where $f_X : \mathbb{R} \to \mathbb{R}$ is the derivative of F_X in the sense of Definition 482.

Definition 505 *Assuming that F_X is absolutely continuous on $\mathbb{R}$, we call the* density *of the real random variable X the equivalence class of all the Lebesgue integrable real functions of real variable satisfying (5.63) of Theorem 501, which are almost everywhere equal. We call a* version of the density *of X any derivative of F_X or equivalently any Lebesgue integrable real function of real variable fulfilling Equation (5.63) of Theorem 501. Hence, a* version of the density *of X is any representative of the equivalence class defining the density of X.*

However, for our purposes it is possible to neglect the distinction between the density of a random variable and a version of the density. Therefore, by the *density* of the real andom variable X we will mean any version of the density of X.

Remark 506 *If f_X is the density of an absolutely continuous real random variable X, we have*

$$\int_{\mathbb{R}} f_X(x)\, d\mu_L(x) = 1.$$

5.2.4 Median of a Real Random Variable

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space, let $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu_L) \equiv \mathbb{R}$ be the Euclidean real line equipped with the Borel σ-algebra $\mathcal{B}(\mathbb{R})$ and the Lebesgue measure μ_L, let $X : \Omega \to \mathbb{R}$ be a real random variable on Ω, and let $F_X : \mathbb{R} \to \mathbb{R}$ be the distribution function of X.

Definition 507 *We call a* median *of X any $x \in \mathbb{R}$ such that*

$$\mathbf{P}(X \leq x) \geq \frac{1}{2} \quad \textit{and} \quad \mathbf{P}(X \geq x) \geq \frac{1}{2}. \tag{5.64}$$

Notation 508 *We write*

$$Q_{1/2} \equiv \left\{ x \in \mathbb{R} : \mathbf{P}(X \leq x) \geq \frac{1}{2} \quad \textit{and} \quad \mathbf{P}(X \geq x) \leq \frac{1}{2} \right\} \tag{5.65}$$

for the set of all $x \in \mathbb{R}$ which are a median of X.

Proposition 509 *A real number x is a median of X if and only if we have*

$$F_X(x) \geq \frac{1}{2} \quad \text{and} \quad \lim_{u \to x^-} F_X(u) \leq \frac{1}{2}. \tag{5.66}$$

Proof. According to the definition of distribution function, we have

$$F_X(x) = \mathbf{P}(X \leq x).$$

Moreover, by virtue of (5.48) in Proposition 473, we know that

$$\lim_{u \to x^-} F_X(u) = \mathbf{P}(X < x) = 1 - \mathbf{P}(X \geq x).$$

The claim clearly follows. □

Example 510 *Given any $x_0 \in \mathbb{R}$, let X be the Dirac real random variable concentrated at x_0 (see Definition 406). Then, X has a unique median at x_0.*

Discussion. Considering the distribution function of the Dirac real random variable concentrated at x_0 (see Example 466) it is clearly seen that we have

$$F_X(x) \geq \frac{1}{2} \Leftrightarrow x \geq x_0.$$

In addition,

$$\lim_{u \to x^-} F_X(u) \leq \frac{1}{2} \Leftrightarrow x \leq x_0.$$

Hence, x_0 is the median for X. □

Example 511 *Given any pair of points $x_0, x_1 \in \mathbb{R}$, such that $x_0 < x_1$, let X be the Bernoulli random variable with states x_0, x_1 and success probability p (see Definition 416). Then, we have*

$$Q_{1/2} = \begin{cases} \{x_0\}, & \text{if } p < 1/2, \\ [x_0, x_1], & \text{if } p = 1/2, \\ \{x_1\}, & \text{if } p > 1/2. \end{cases} \tag{5.67}$$

Discussion. The distribution function of the Bernoulli random variable with states x_0, x_1 and success probability p is given by

$$F_X(x) = \begin{cases} 0, & \text{if } x < x_0, \\ 1-p, & \text{if } x_0 \leq x < x_1, \\ 1, & \text{if } x_1 \leq x, \end{cases}$$

(see Example 468). Hence, under the assumption $p < 1/2$, that is $1 - p > 1/2$, we have

$$F_X(x) \geq \frac{1}{2} \Leftrightarrow F_X(x) \geq 1 - p \Leftrightarrow x \geq x_0 \tag{5.68}$$

and

$$\lim_{u \to x^-} F_X(u) \leq \frac{1}{2} \Leftrightarrow \lim_{u \to x^-} F_X(u) \leq 0 \Leftrightarrow x \leq x_0. \tag{5.69}$$

Combining (5.68) and (5.69), we obtain $Q_{1/2} = \{x_0\}$.

Now, assume that $p = 1/2$, that is $1 - p = 1/2$. Then, Equation (5.68) still holds true and we have

$$\lim_{u \to x^-} F_X(u) \leq \frac{1}{2} \Leftrightarrow \lim_{u \to x^-} F_X(u) \leq 1 - p \Leftrightarrow x \leq x_1. \tag{5.70}$$

By virtue of (5.68) and (5.68), it Then, follows that $Q_{1/2} = [x_0, x_1]$.

In the end, if $p > 1/2$, that is $1 - p < 1/2$, we have

$$F_X(x) \geq \frac{1}{2} \Leftrightarrow F_X(x) > 1 - p \Leftrightarrow x \geq x_1 \tag{5.71}$$

and

$$\lim_{u \to x^-} F_X(u) \leq \frac{1}{2} \Leftrightarrow \lim_{u \to x^-} F_X(u) \leq 1 - p \Leftrightarrow x \leq x_1. \tag{5.72}$$

Combining (5.71) and (5.72), we obtain $Q_{1/2} = \{x_1\}$. □

Proposition 512 *If the real random variable X is symmetric about x_0, Then, it has a median at x_0.*

Proof. As a consequence of Equation (5.36), if X is symmetric about x_0 we have

$$\mathbf{P}(X \leq x_0) = \mathbf{P}(X \geq x_0).$$

On the other hand,

$$\mathbf{P}(X \leq x_0) + \mathbf{P}(X \geq x_0) \geq 1.$$

Combining the above equations, it immediately follows that x_0 fulfills Inequalities (5.64) characterizing a median for X. □

Proposition 513 *The set $Q_{1/2}$ is always non-empty. In other words: any real random variable has at least a median.*

Proof. Consider the set

$$\hat{Q}_{1/2} \equiv \left\{ x \in \mathbb{R} : F_X(x) \geq \frac{1}{2} \right\}.$$

Since $\lim_{x \to +\infty} F_X(x) = 1$, we have that $\hat{Q}_{1/2} \neq \varnothing$. Moreover, since $\lim_{x \to -\infty} F_X(x) = 0$ and the distribution function F_X is non decreasing, we have that $\hat{Q}_{1/2}$ has a lower bound. Write

$$\hat{x}_{1/2} \equiv \inf \hat{Q}_{1/2}. \tag{5.73}$$

We have

$$F_X\left(\hat{x}_{1/2}\right) \geq \frac{1}{2}. \tag{5.74}$$

In fact, if we had $F_X\left(\hat{x}_{1/2}\right) < \frac{1}{2}$, since $\lim_{x \to \hat{x}_{1/2}^+} F_X(x) = F_X\left(\hat{x}_{1/2}\right)$, there would exist $\delta > 0$ such that $F_X(x) < \frac{1}{2}$ for every $x \in \left[\hat{x}_{1/2}, \hat{x}_{1/2} + \delta\right)$. On the other hand, the function F_X is not decreasing on $\mathbb{R}$. Hence,

$$F_X(x) \leq F_X\left(\hat{x}_{1/2}\right) < \frac{1}{2},$$

for every $x \leq \hat{x}_{1/2}$. It would follow

$$\left(-\infty, \hat{x}_{1/2} + \delta\right) \cap \hat{Q}_{1/2} = \varnothing,$$

which would prevent that $\hat{x}_{1/2} \equiv \inf \hat{Q}_{1/2}$.We have also

$$\lim_{x \to \hat{x}_{1/2}^-} F_X(x) \leq \frac{1}{2}. \tag{5.75}$$

In fact, if we had $\lim_{x \to \hat{x}_{1/2}^-} F_X(x) > \frac{1}{2}$, there would exist $\delta > 0$ such that $F_X(x) > \frac{1}{2}$ for every $x \in \left(\hat{x}_{1/2} - \delta, \hat{x}_{1/2}\right)$. Hence, we would have

$$\left(\hat{x}_{1/2} - \delta, \hat{x}_{1/2}\right) \subseteq \hat{Q}_{1/2},$$

which would prevent again that $\hat{x}_{1/2} \equiv \inf \hat{Q}_{1/2}$. In the end, having proved that Equations (5.74) and (5.75) hold true, we have shown the existence of a median. A similar proof is based on considering the set

$$\check{Q}_{1/2} \equiv \left\{ x \in \mathbb{R} : \lim_{u \to x^-} F_X(u) \leq \frac{1}{2} \right\}.$$

Since $\lim_{x \to -\infty} F_X(x) = 0$, we have that $\hat{Q}_{1/2} \neq \varnothing$. Moreover, since $\lim_{x \to +\infty} F_X(x) = 1$ and the distribution function F_X is non decreasing, the set $\check{Q}_{1/2}$ has an upper bound. Write

$$\check{x}_{1/2} \equiv \sup \check{Q}_{1/2}. \tag{5.76}$$

We have

$$\lim_{x \to \check{x}_{1/2}^-} F_X(x) \leq \frac{1}{2}. \tag{5.77}$$

In fact, if we had $\lim_{x \to \check{x}_{1/2}^-} F_X(x) > \frac{1}{2}$, there would exist $\delta > 0$ such that $F_X(x) > \frac{1}{2}$ for every $x \in \left(\check{x}_{1/2} - \delta, \check{x}_{1/2}\right)$. On the other hand, the distribution function F_X is not decreasing on $\mathbb{R}$. Hence,

$$F_X(x) \geq F_X\left(\check{x}_{1/2}\right) \geq \lim_{x \to \check{x}_{1/2}^-} F_X(x) > \frac{1}{2},$$

for every $x \geq \check{x}_{1/2}$. It would follow

$$\left(\check{x}_{1/2} - \delta, +\infty\right) \cap \hat{Q}_{1/2} = \varnothing,$$

which would prevent that $\check{x}_{1/2} \equiv \sup \check{Q}_{1/2}$. We have also

$$F_X\left(\check{x}_{1/2}\right) \geq \frac{1}{2}. \tag{5.78}$$

In fact, if we had $F_X\left(\hat{x}_{1/2}\right) < \frac{1}{2}$, since $\lim_{x \to \check{x}_{1/2}^+} F_X(x) = F_X\left(\check{x}_{1/2}\right)$, there would exist $\delta > 0$ such that $F_X(x) < \frac{1}{2}$ for every $x \in \left[\check{x}_{1/2}, \check{x}_{1/2} + \delta\right)$. As a consequence,

$$\lim_{u \to x^-} F_X(u) \leq \frac{1}{2}$$

for every $x \in \left[\check{x}_{1/2}, \check{x}_{1/2} + \delta\right)$. This would imply that

$$\left[\check{x}_{1/2}, \check{x}_{1/2} + \delta\right) \subseteq \check{Q}_{1/2},$$

which would prevent again that $\check{x}_{1/2} \equiv \sup \check{Q}_{1/2}$. Thus, having proved that Equations (5.77) and (5.78) hold true, we have shown again the existence of a median. □

Remark 514 *With reference to the points $\hat{x}_{1/2}$ and $\check{x}_{1/2}$, introduced by Equations (5.73) and (5.76) in the Proof of Proposition 513, we have*

$$\hat{x}_{1/2} = \min \hat{Q}_{1/2} \quad \textit{and} \quad \check{x}_{1/2} = \max \check{Q}_{1/2}. \tag{5.79}$$

Definition 515 *The point $\hat{x}_{1/2}$ [resp. $\check{x}_{1/2}$] introduced by Equation (5.73) [resp. (5.76)] in the Proof of Proposition 513 is called* the minimum *[resp.* maximum*]* median *of X.*

Proposition 516 *Let $\hat{x}_{1/2}$ and $\check{x}_{1/2}$ be the minimum and maximum median of X, respectively. Then, we have*

$$Q_{1/2} = \left[\hat{x}_{1/2}, \check{x}_{1/2}\right]. \tag{5.80}$$

Proof. According to the definition of $\hat{x}_{1/2}$ and $\check{x}_{1/2}$, we clearly have

$$Q_{1/2} \subseteq \left[\hat{x}_{1/2}, \check{x}_{1/2}\right]. \tag{5.81}$$

On the other hand, since the function F_X is not decreasing on $\mathbb{R}$, we have

$$F_X(x) \geq F_X\left(\hat{x}_{1/2}\right) \geq \frac{1}{2}, \tag{5.82}$$

for every $x \geq \hat{x}_{1/2}$ and

$$\lim_{u \to x^-} F_X(u) \leq \lim_{x \to \check{x}_{1/2}^-} F_X(x) \leq \frac{1}{2}, \tag{5.83}$$

for every $x \leq \check{x}_{1/2}$. From (5.114) and (5.115) we obtain

$$\left[\hat{x}_{1/2}, \check{x}_{1/2}\right] \subseteq Q_{1/2}. \tag{5.84}$$

Combining (5.81) and (5.84) the desired (5.80) clearly follows. □

Definition 517 *In case the random variable X has a unique median, we call it* the median *of X and denote it by $x_{1/2}$.*

Proposition 518 *If the real random variable X has strictly increasing distribution function F_X, Then, a median of X is unique.*

Proof. Thanks to (5.80), we know that

$$Q_{1/2} = \left[\hat{x}_{1/2}, \check{x}_{1/2}\right].$$

In particular,

$$F_X\left(\hat{x}_{1/2}\right) \geq \frac{1}{2} \quad \text{and} \quad \lim_{x \to \check{x}_{1/2}^-} F_X(x) \leq \frac{1}{2}. \tag{5.85}$$

Now, when F_X is strictly increasing, if we had $\hat{x}_{1/2} \neq \check{x}_{1/2}$, it would follow

$$F_X\left(\hat{x}_{1/2}\right) < \lim_{x \to \check{x}_{1/2}^-} F_X(x). \tag{5.86}$$

From (5.85) and (5.86) we would obtain a clear contradiction. □

Proposition 519 *If the real random variable X has strictly increasing and continuous distribution function F_X, Then, a median of X is unique and fulfills*

$$F_X\left(x_{1/2}\right) = \frac{1}{2}. \tag{5.87}$$

Proof. On account of 1 and 2 in Proposition 472, the continuity of F_X implies that there exists at least $x_{1/2} \in \mathbb{R}$ such that

$$F_X\left(x_{1/2}\right) = 1/2. \tag{5.88}$$

Still the continuity of F_X implies that

$$\lim_{u \to x_{1/2}^-} F_X(u) = F_X\left(x_{1/2}\right) = 1/2. \tag{5.89}$$

From (5.88) and (5.89) it clearly follows that $x_{1/2}$ is a median of X. In the end, since F_X is strictly increasing and continuous, we have

$$\lim_{u \to x^-} F_X(u) = F_X(x) < 1/2 \quad \text{and} \quad \lim_{u \to x^-} F_X(u) = F_X(x) > 1/2$$

according to whether $x < x_{1/2}$ or $x > x_{1/2}$. This prevents Equation (5.66) of Proposition 509 from being fulfilled for any $x \in \mathbb{R}$ other than $x_{1/2}$, that is the uniqueness of $x_{1/2}$. □

Corollary 520 *If the real random variable X is absolutely continuous and has strictly positive density $f_X : \mathbb{R} \to \mathbb{R}$, Then, a median is unique and fulfills*

$$\int_{\left(-\infty, x_{1/2}\right]} f_X(x)\, d\mu_L(x) = \frac{1}{2}. \tag{5.90}$$

Proof. The absolute continuity of the random variable X means the absolute continuity of the distribution function F_X which fulfills

$$F_X(x) = \int_{(-\infty, x]} f_X(u)\, d\mu_L(u), \tag{5.91}$$

for every $x \in \mathbb{R}$. In addition, since f_X is strictly positive F_X is strictly increasing. Hence, by virtue of Proposition 519, there exists a unique median $x_{1/2}$ such that

$$F_X\left(x_{1/2}\right) = 1/2 \tag{5.92}$$

Combining (5.91) and (5.92), the desired (5.90) follows. □

5.2.5 Quantiles of a Real Random Variable

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space, let $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu_L) \equiv \mathbb{R}$ be the Euclidean real line equipped with the Borel σ-algebra $\mathcal{B}(\mathbb{R})$ and the Lebesgue measure μ_L, let $X : \Omega \to \mathbb{R}$ be a real random variable on Ω, and let $F_X : \mathbb{R} \to \mathbb{R}$ be the distribution function of X.

Definition 521 *Given any $q \in (0, 1)$, we call a* quantile of order q *or* q-quantile *of X any $x \in \mathbb{R}$ such that*

$$\mathbf{P}(X \leq x) \geq q \quad \textit{and} \quad \mathbf{P}(X \geq x) \geq 1 - q. \tag{5.93}$$

Remark 522 *A quantile of order* $1/2$ *of* X *is just a median.*

Definition 523 *With regard to a more specific nomenclature, we call:*

1. tercile *or* tertile *any* $k/3$*-quantile for* $k = 1, 2$*; in particular, on varying of* $k = 1, 2$*, the* $k/3$*-tercile is called the* k*th tercile;*
2. quartile *any* $k/4$*-quantile for* $k = 1, 2, 3$*; in particular, on varying of* $k = 1, 2, 3$*, the* $k/4$*-quartile is called the* k*th quartile;*
3. quintile *any* $k/5$*-quantile for* $k = 1, \ldots, 4$*; in particular, on varying of* $k = 1, \ldots, 4$*, the* $k/5$*-quintile is called the* k*th quintile;*
4. sextile *any* $k/6$*-quantile for* $k = 1, \ldots, 5$*; in particular, on varying of* $k = 1, \ldots, 5$*, the* $k/6$*-sextile is called the* k*th sextile;*
5. septile *any* $k/7$*-quantile for* $k = 1, \ldots, 6$*; in particular, on varying of* $k = 1, \ldots, 6$*, the* $k/7$*-septile is called the* k*th septile;*
6. octile *any* $k/8$*-quantile for* $k = 1, \ldots, 7$*; in particular, on varying of* $k = 1, \ldots, 7$*, the* $k/8$*-octile is called the* k*th octile;*
7. decile *any* $k/10$*-quantile for* $k = 1, \ldots, 9$*; in particular, on varying of* $k = 1, \ldots, 9$*, the* $k/9$*-decile is called the* k*th decile;*
8. duo-decile *any* $k/12$*-quantile for* $k = 1, \ldots, 11$*; in particular, on varying of* $k = 1, \ldots, 11$*, the* $k/12$*-duo-decile is called the* k*th duo-decile;*
9. hexadecile *any* $k/16$*-quantile for* $k = 1, \ldots, 15$*; in particular, on varying of* $k = 1, \ldots, 15$*, the* $k/15$*-hexadecile is called the* k*th hexadecile;*
10. ventile *or* vigintile *any* $k/20$*-quantile for* $k = 1, \ldots, 19$*; in particular, on varying of* $k = 1, \ldots, 19$*, the* $k/19$*-ventile is called the* k*th ventile;*
11. centile *or* percentile *any* $k/100$*-quantile for* $k = 1, \ldots, 99$*; in particular, on varying of* $k = 1, \ldots, 99$*, the* $k/100$*-centile is called the* k*th centile;*
12. permille *any* $k/1000$*-quantile for* $k = 1, \ldots, 999$*; in particular, on varying of* $k = 1, \ldots, 999$*, the* $k/1000$*-permille is called the* k*th permille.*

A similar nomenclature is presented in the following Definition.

Definition 524 *Given* $n \geq 2$*, any* k/n*-quantile for* $k = 1, \ldots, n-1$ *is called* n-ile*; in particular, for any* $m \in \{1, \ldots, n\}$*, the* m/n*-quantile of order is called* the m-th n-ile.

Note also that some authors adopt a slightly different nomenclature.

Definition 525 *For any* $q \in (0, 1)$ *the* q-quantile *of* X *is also called the* $(100q)$*th percentile of* X*.*

For instance, according to this nomenclature, the median of X is also called the 50th percentile of X. Eventually, for $q = k/100$, where $k = 1, \ldots, 99$, the nomenclature in Definition 525 agrees with that of item 11. in Definition 523.

Notation 526 *Given any $q \in (0,1)$, we write*

$$Q_q \equiv \{x \in \mathbb{R} : \mathbf{P}(X \leq x) \geq q \quad \textit{and} \quad \mathbf{P}(X \geq x) \leq 1-q\} \tag{5.94}$$

for the set of all $x \in \mathbb{R}$ which are a q-quantile of X.

Remark 527 *Given any $q \in (0,1)$, a real number x is a q-quantile of X if and only if any of the following equivalent conditions is fulfilled*

$$\mathbf{P}(X \leq x) \geq q \quad \textit{and} \quad \mathbf{P}(X < x) \leq q, \tag{5.95}$$
$$\mathbf{P}(X > x) \leq 1-q \quad \textit{and} \quad \mathbf{P}(X \geq x) \geq 1-q, \tag{5.96}$$
$$\mathbf{P}(X > x) \leq 1-q \quad \textit{and} \quad \mathbf{P}(X < x) \leq q. \tag{5.97}$$

Proposition 528 *Given any $q \in (0,1)$, a real number x is a q-quantile of X if and only we have*

$$F_X(x) \geq q \quad \textit{and} \quad \lim_{u \to x^-} F_X(u) \leq q. \tag{5.98}$$

Proof. Observing that

$$\mathbf{P}(X \leq x) = F_X(x) \quad \text{and} \quad \mathbf{P}(X < x) = \lim_{x \to x^-} F_X(x),$$

the claim follows from Equation (5.95). □

Example 529 *Given any $x_0 \in \mathbb{R}$, let X be the Dirac real random variable concentrated at x_0 (see Definition 406). Then, X has a unique q-quantile at x_0 for every $q \in (0,1)$*

Discussion. Considering the distribution function of the Dirac real random variable concentrated at x_0 (see Example 466) it is clearly seen that we have

$$F_X(x) \geq q \Leftrightarrow x \geq x_0,$$

for every $q \in (0,1)$. In addition,

$$\lim_{u \to x^-} F_X(u) \leq q \Leftrightarrow x \leq x_0,$$

for every $q \in (0,1)$. Hence, x_0 is the q-quantile of X, for every $q \in (0,1)$. □

Example 530 *Given any pair of points $x_0, x_1 \in \mathbb{R}$, such that $x_0 < x_1$, let X be the Bernoulli random variable with states x_0, x_1 and success probability p (see Definition 416). Then, we have*

$$Q_q = \begin{cases} \{x_0\} & \textit{if } q < 1-p \\ [x_0, x_1] & \textit{if } q = 1-p \\ \{x_1\} & \textit{if } q > 1-p \end{cases}. \tag{5.99}$$

Discussion. The distribution function of the Bernoulli random variable with states x_0, x_1 and success probability p is given by

$$F_X(x) = \begin{cases} 0 & \text{if } x < x_0 \\ 1-p & \text{if } x_0 \leq x < x_1 \\ 1 & \text{if } x_1 \leq x \end{cases}$$

(see Example 468). Hence, under the assumption $q < 1 - p$, we have

$$F_X(x) \geq q \Leftrightarrow F_X(x) \geq 1 - p \Leftrightarrow x \geq x_0 \tag{5.100}$$

and

$$\lim_{u \to x^-} F_X(u) \leq q \Leftrightarrow \lim_{u \to x^-} F_X(u) \leq 0 \Leftrightarrow x \leq x_0. \tag{5.101}$$

Combining (5.68) and (5.69), we obtain $Q_{1/2} = \{x_0\}$.

Now, assume that $q = 1 - p$. Then, Equation (5.100) still holds true and we have

$$\lim_{u \to x^-} F_X(u) \leq q \Leftrightarrow \lim_{u \to x^-} F_X(u) \leq 1 - p \Leftrightarrow x \leq x_1. \tag{5.102}$$

By virtue of (5.68) and (5.68), it Then, follows that $Q_q = [x_0, x_1]$.

In the end, if $q > 1 - p$, we have

$$F_X(x) \geq q \Leftrightarrow F_X(x) \geq 1 - p \Leftrightarrow x \geq x_1 \tag{5.103}$$

and

$$\lim_{u \to x^-} F_X(u) \leq q \Leftrightarrow \lim_{u \to x^-} F_X(u) \leq 1 - p \Leftrightarrow x \leq x_1. \tag{5.104}$$

Combining (5.103) and (5.104), we obtain $Q_q = \{x_1\}$. □

Proposition 531 *Given any $q \in (0,1)$, the set Q_q is always non-empty. In other words: any real random variable has at least a q-quantile for every $q \in (0,1)$.*

Proof. The proof follows very closely that of Proposition 513. Given any $q \in (0,1)$, consider the set

$$\hat{Q}_q \equiv \{x \in \mathbb{R} : F_X(x) \geq q\}.$$

Since $\lim_{x \to +\infty} F_X(x) = 1$, we have that $\hat{Q}_q \neq \varnothing$. Moreover, since $\lim_{x \to -\infty} F_X(x) = 0$ and the distribution function F_X is non decreasing, we have that $\hat{Q}_{1/2}$ has a lower bound. Write

$$\hat{x}_q \equiv \inf \hat{Q}_q. \tag{5.105}$$

We have

$$F_X(\hat{x}_q) \geq q. \tag{5.106}$$

In fact, if we had $F_X(\hat{x}_q) < q$, since $\lim_{x \to \hat{x}_q^+} F_X(x) = F_X(\hat{x}_q)$, there would exist $\delta > 0$ such that $F_X(x) < q$ for every $x \in [\hat{x}_q, \hat{x}_q + \delta)$. On the other hand, the function F_X is not decreasing on $\mathbb{R}$. Hence,

$$F_X(x) \leq F_X(\hat{x}_q) < q$$

for every $x \leq \hat{x}_q$. It would follow

$$(-\infty, \hat{x}_q + \delta) \cap \hat{Q}_q = \varnothing,$$

which would prevent that $\hat{x}_q \equiv \inf \hat{Q}_{1/2}$.We have also

$$\lim_{x \to \hat{x}_q^-} F_X(x) \leq q. \tag{5.107}$$

In fact, if we had $\lim_{x\to\hat{x}_q^-} F_X(x) > q$, there would exist $\delta > 0$ such that $F_X(x) > q$ for every $x \in (\hat{x}_q - \delta, \hat{x}_q)$. It would follow

$$(\hat{x}_q - \delta, \hat{x}_q) \subseteq \hat{Q}_q,$$

which would prevent again that $\hat{x} \equiv \inf \hat{Q}_q$. In the end, having proved that Equations (5.106) and (5.107) hold true, we have shown the existence of a quantile of order q. Still as in the proof of Proposition 513 we could consider the set

$$\check{Q}_q \equiv \left\{ x \in \mathbb{R} : \lim_{u\to x^-} F_X(u) \leq q \right\}.$$

Since $\lim_{x\to-\infty} F_X(x) = 0$, we have that $\hat{Q}_q \neq \varnothing$. Moreover, since $\lim_{x\to+\infty} F_X(x) = 1$ and the distribution function F_X is non decreasing, the set $\check{Q}_{1/2}$ has an upper bound. Write

$$\check{x}_q \equiv \sup \check{Q}_q. \tag{5.108}$$

We have

$$\lim_{x\to\check{x}_q^-} F_X(x) \leq q. \tag{5.109}$$

In fact, if we had $\lim_{x\to\check{x}_q^-} F_X(x) > \frac{1}{2}$, there would exist $\delta > 0$ such that $F_X(x) > q$ for every $x \in (\check{x}_q - \delta, \check{x}_q)$. On the other hand, the distribution function F_X is not decreasing on $\mathbb{R}$. Hence,

$$F_X(x) \geq F_X(\check{x}_q) \geq \lim_{x\to\check{x}_q^-} F_X(x) > q$$

for every $x \geq \check{x}_q$. It would follow

$$(\check{x}_q - \delta, +\infty) \cap \check{Q}_q = \varnothing,$$

which would prevent that $\check{x}_q \equiv \sup \check{Q}_q$. We have also

$$F_X(\check{x}_q) \geq q. \tag{5.110}$$

In fact, if we had $F_X(\check{x}_q) < q$, since $\lim_{x\to\check{x}_q^+} F_X(x) = F_X(\check{x}_q)$, there would exist $\delta > 0$ such that $F_X(x) < q$ for every $x \in [\check{x}_q, \check{x}_q + \delta)$. As a consequence,

$$\lim_{u\to x^-} F_X(u) \leq q,$$

for every $x \in [\check{x}_q, \check{x}_q + \delta)$. This would imply that

$$[\check{x}_q, \check{x}_q + \delta) \subseteq \check{Q}_q,$$

which would prevent again that $\check{x}_q \equiv \sup \check{Q}_q$. Thus, having proved that Equations (5.109) and (5.110) hold true, we have shown again the existence of a quantile of order q. □

Remark 532 *With reference to the points $\hat{x}_q$ and $\check{x}_q$, introduced by Equations (5.105) and (5.108) in the Proof of Proposition 531, we have*

$$\hat{x}_q = \min \hat{Q}_q \quad \textit{and} \quad \check{x}_q = \max \check{Q}_q. \tag{5.111}$$

Definition 533 *Given any $q \in (0,1)$, the point $\hat{x}_q$ [resp. $\check{x}_q$] introduced by Equation (5.105) [resp. (5.108)] in the Proof of Proposition 531 is called* the minimum *[resp.* maximum*]* q-quantile *of X.*

Proposition 534 *Given any $q \in (0,1)$, let $\hat{x}_q$ and $\check{x}_q$ be the minimum and maximum q-*quantile *of X, respectively. Then, we have*

$$Q_q = [\hat{x}_q, \check{x}_q]. \tag{5.112}$$

Proof. The proof is just a reformulation of that of Proposition 516. According to the definition of $\hat{x}$ and $\check{x}$, we clearly have

$$Q_{1/2} \subseteq [\hat{x}_q, \check{x}_q]. \tag{5.113}$$

On the other hand, since the function F_X is not decreasing on $\mathbb{R}$, we have

$$F_X(x) \geq F_X(\hat{x}_q) \geq q, \tag{5.114}$$

for every $x \geq \hat{x}_q$ and

$$\lim_{u \to x^-} F_X(u) \leq \lim_{x \to \check{x}_q} F_X(x) \leq q, \tag{5.115}$$

for every $x \leq \check{x}_q$. From (5.114) and (5.115) we obtain

$$[\hat{x}_q, \check{x}_q] \subseteq Q_q. \tag{5.116}$$

Hence, combing (5.113) and (5.116) the desired (5.112) clearly follows. □

Definition 535 *In case the random variable X has a unique q-quantile, for some $q \in (0,1)$, we call it* the q-quantile *of X and denote it by x_q.*

Definition 536 *We call* quantile function *of X the function $Q_X : (0,1) \to \mathbb{R}$ given by*

$$Q_X(q) \stackrel{def}{=} \hat{x}_q,$$

where $\hat{x}_q$ is the minimum q-quantile of X.

Proposition 537 *We have:*

1. *the function Q_X is increasing on $(0,1)$;*
2. *$Q_X(F_X(x)) \leq x$ for every $x \in \mathbb{R}$ such that $0 < F_X(x) < 1$;*
3. *$F_X(Q_X(q)) \geq q$ for every $q \in (0,1)$;*
4. *Q_X is left-continuous, that is $\lim_{u \to q^-} Q_X(u) = Q_X(q)$ for every $q \in (0,1)$;*
5. *Q_X has finite right-limits, more specifically $\lim_{v \to q^+} Q_X(v) = \inf\{x \in \mathbb{R} : F_X(x) > q\}$ for every $q \in (0,1)$.*

Proof. . □

Proposition 538 *If the real random variable X has strictly increasing distribution function F_X, Then, a q-quantile of X is unique for every $q \in (0,1)$.*

Proof. We just rewrite the Proof of Proposition 518. Thanks to (5.112) in Proposition 534, we know that

$$Q_q = [\hat{x}_q, \check{x}_q] .$$

dor every $q \in (0,1)$. In particular,

$$F_X(\hat{x}_q) \geq q \quad \text{and} \quad \lim_{x \to \check{x}_q^-} F_X(x) \leq q. \tag{5.117}$$

Now, when F_X is strictly increasing, if we had $\hat{x}_q \neq \check{x}_q$, it would follow

$$F_X(\hat{x}_q) < \lim_{x \to \check{x}_q^-} F_X(x) . \tag{5.118}$$

From (5.117) and (5.118) we would obtain a clear contradiction. □

Corollary 539 *If the real random variable X has strictly increasing distribution function F_X, Then, we have*

$$Q_X = F_X^{-1}.$$

Proposition 540 *If the real random variable X has strictly increasing and continuous distribution function F_X, Then, a q-quantile of X is unique and fulfills*

$$F_X(x_q) = q. \tag{5.119}$$

for every $q \in (0,1)$.

Proof. In this case, the proof follows that of Proposition 519. On account of 1 and 2 of Proposition 472, the continuity of F_X implies that there exists at least $x_q \in \mathbb{R}$ such that

$$F_X(x_q) = q \tag{5.120}$$

for every $q \in (0,1)$. Still the continuity of F_X implies that

$$\lim_{u \to x_q^-} F_X(u) = F_X(x_q) = q. \tag{5.121}$$

From (5.120) and (5.121) it clearly follows that x_q is a quantile of X. In the end, since F_X is strictly increasing and continuous, we have

$$\lim_{u \to x^-} F_X(u) = F_X(x) < q \quad \text{and} \quad \lim_{u \to x^-} F_X(u) = F_X(x) > q$$

according to whether $x < x_q$ or $x > x_q$. This prevents Equation (5.98) of Proposition 528 from being fulfilled for any $x \in \mathbb{R}$ other than x_q, that is the uniqueness of x_q. □

Corollary 541 *If the real random variable X is absolutely continuous and has strictly positive density $f_X : \mathbb{R} \to \mathbb{R}$, Then, a q-quantile of X is unique and fulfills*

$$\int_{(-\infty, x_q]} f_X(x)\, dx = q, \tag{5.122}$$

for every $q \in (0,1)$.

Example 542 *Assume X is standard continuous uniformly distributed. Then, we have*

$$x_q = q, \tag{5.123}$$

for every $q \in (0,1)$.

Discussion. By virtue of Equation (5.180), we know that

$$F_X(x) = x1_{[0,1]}(x) + 1_{(1,+\infty)}(x), \tag{5.124}$$

for every $x \in \mathbb{R}$. Now, when $q \in (0,1)$ the condition

$$F_X(x) \geq q \quad \text{and} \quad \lim_{u \to x^-} F_X(u) \leq q \tag{5.125}$$

implies that $x \in (0,1)$. Therefore, since the function F_X is strictly increasing and continuous in $(0,1)$, Condition (5.124) becomes

$$F_X(x) = q$$

Still on account of (5.124), the latter yields

$$x = q.$$

as desired. □

Q-Q plots

A QQ plot, where QQ stands for quantile-quantile, is a graphical method in the Cartesian plane $\mathbb{R}^2$ to compare two probability distributions by plotting their quantiles against each other.

Ler X, Y real random variables with strictly increasing and continuous distribution function $F_X : \mathbb{R} \to \mathbb{R}$ and $F_Y : \mathbb{R} \to \mathbb{R}$, respectively.

Definition 543 *We call the* Q-Q plot of Y against X*, the representation in the Cartesian plane $\mathbb{R}^2$ of the parametric curve $QQ_{X,Y} : (0,1) \to \mathbb{R}^2$ given by*

$$QQ_{X,Y} \stackrel{def}{=} (x_q, y_q) \quad \forall q \in (0,1),$$

where x_q and y_q fulfill the equations

$$F_X(x_q) = F_Y(y_q) = q.$$

Let $(x_k)_{k=1}^N \equiv x$ and $(y_k)_{k=1}^N \equiv y$ two data set of the same size N and let $\left(x_{(k)}\right)_{k=1}^N \equiv x_{()}$ and $\left(y_{(k)}\right)_{k=1}^N \equiv y_{()}$ be the order statistics of x and y, respectively. We agree that x [resp. y] is a set of data generated from a reference [resp. tested] probability distribution.

Definition 544 *We call the data set x [resp. y] the* reference *[resp.* test*] data set.*

Definition 545 *We call the* Q-Q plot of y against x*, the representation in the Cartesian plane $\mathbb{R}^2$ of the parametric point curve $QQ_{x,y} : \{1, \ldots, N\} \to \mathbb{R}^2$ given by*

$$QQ_{x,y}(k) \stackrel{def}{=} (x_k, y_k) \quad \forall k \in \{1, \ldots, N\}.$$

Analogously to the case of a PP plot, the shape of the pattern of a Q-Q plot provides us with several information about the probability distributions which generates the reference and test data.

Remark 546 *Assume $Y = X$. Then, the pattern of the corresponding Q-Q plot $QQ_{X,Y}$ lies on the straigth line $y = x$. Conversely, assume that the pattern of the Q-Q plot $QQ_{X,Y}$ lies on the straigth line $y = x$. Then, $Y = X$.*

Remark 547 *Assume the test data set y is drawn from the same distribution generating the reference data set x. Then, the pattern of the corresponding Q-Q plot $QQ_{x,y}$ is very close to the straigth line $y = x$. Conversely, assume that the pattern of the Q-Q plot $QQ_{x,y}$ is very close to the straigth line $y = x$, Then, the test data set y is likely drawn from the same distribution which generates the reference data set x.*

Remark 548 *Assume $Y = aX + b$, where $a, b \in \mathbb{R}$ with $a \neq 0$. Then, the pattern of the corresponding Q-Q plot $QQ_{X,Y}$ lies on the straigth line $y = ax + b$. Conversely, assume that the pattern of the Q-Q plot $QQ_{X,Y}$ lies on the straigth line $y = ax + b$, Then, $Y = aX + b$.*

Remark 549 *Assume the test data set y is drawn from a distribution which is a linear transformation of the distribution generating the reference data set x. Then, the pattern of the corresponding Q-Q plot $QQ_{x,y}$ is very close to a straight line. Conversely, assume that the pattern of the Q-Q plot $QQ_{x,y}$ is very close to a straight line, Then, the test data set y is likely drawn from a linear transformation of the distribution which generates the reference data set x.*

Remark 550 *Assume X and Y belong to the same family of distribution but it is not true that $Y = aX + b$, for some $a, b \in \mathbb{R}$ with $a \neq 0$. Then, the differences in the characterizing parameters of X and Y generally do not allow a straight line shape of the pattern of the Q-Q plot $QQ_{X,Y}$.*

Remark 551 *Assume the data sets x and y are drawn from the same family of distribution but it is not true that the distribution which generates the test data set y is a linear transformation of the distribution generating the reference data set x. Then, the differences in the characterizing parameters of the distributions which generate x and y do not generally allow a straight line shape of the pattern of the Q-Q plot $QQ_{x,y}$.*

Remark 552 *Assume the distribution of Y is more concentrated [resp. dispersed] than the distribution of X. Then, pattern of the the corresponding Q-Q plot $QQ_{X,Y}$ is flatter [resp. steeper] than the straigth line $y = x$. Conversely, assume that the pattern of the Q-Q plot $QQ_{X,Y}$ is flatter [resp. steeper] than the straigth line $y = x$. Then, the distribution of Y is more concentrated [resp. dispersed] than the distribution of X.*

Remark 553 *Assume the test data set y is drawn from a distribution which is more concentrated [resp. dispersed] than the distribution generating the reference data set x. Then, the pattern of the corresponding Q-Q plot $QQ_{x,y}$ is flatter [resp. steeper] than the straigth line $y = x$. Conversely, assume that the pattern of the Q-Q plot $QQ_{x,y}$ is flatter [resp. steeper] than the straigth line $y = x$. Then, the test data set y is likely drawn from a distribution which is more concentrated [resp. dispersed] than the distribution generating the reference data set x.*

Remark 554 *Assume the distribution of Y is more concentrated on the left [resp. right] than the distribution of X. Then, the pattern of the corresponding Q-Q plot $QQ_{X,Y}$ is arched downwards [resp. upwards]. Conversely, assume that the pattern of the Q-Q plot $QQ_{X,Y}$ is arched downwards [resp. upwards]. Then, the distribution of Y is more concentrated on the left [resp. right] than the distribution of X.*

Remark 555 *Assume the test data set y is drawn from a distribution which is more concentrated on the left [resp. right] than the distribution of X. Then, the pattern of the corresponding Q-Q plot $QQ_{X,y}$ is arched downwards [resp. upwards]. Conversely, assume that the pattern of the Q-Q plot $QQ_{X,y}$ arched downwards [resp. upwards]. Then, the test data set y is likely drawn from a distribution which is more concentrated on the left [resp. right] than the distribution of X.*

Remark 556 *Assume the distribution of Y has lighter [resp. heavier] tails than the distribution of X. Then, the pattern of the corresponding Q-Q plot $QQ_{X,Y}$ is "S" [resp. reverse "S"]-shaped like the logistic [resp. logit]. Conversely, assume that the pattern of the Q-Q plot $QQ_{X,Y}$ is "S" [resp. reverse "S"]-shaped like the logistic [resp. logit]. Then, the distribution of Y has lighter [resp. heavier] tails than the distribution of X.*

Remark 557 *Assume the test data set y is drawn from a distribution which has lighter [resp. heavier] tails than the distribution of X. Then, the pattern of the corresponding Q-Q plot $QQ_{X,y}$ is "S" [resp. reverse "S"]-shaped like the logistic [resp. logit]. Conversely, assume that the pattern of the Q-Q plot $QQ_{X,y}$ is "S" [resp. reverse "S"]-shaped like the logistic [resp. logit]. Then, the test data set y is likely drawn from a distribution which has lighter [resp. heavier] tails than the distribution of X.*

Note that exchanging the test and the reference data sets results in a reversion of the shape of the associated Q-Q plot. In general a Q-Q plot is more sensitive to deviances from normality in the tails of the data set than the corresponding P-P plot, whereas a P-P plot is more sensitive to deviances near the mean of the distribution than the corresponding Q-Q plot. This makes a Q-Q plot a better detector of outliers in the test data set that a P-P plot. The identification of outliers in a data set is an important goal. Therefore, Q-Q plots are more frequently used than P-P plots to assess the normality of a data set.

5.2.6 More about the median and the quantiles of a random variable

In this subsection, we consider some additional issue about the median and the quantiles of a random variable.

Example 558 *Let $X \tilde{} Bin(n,p)$ be the standard binomial random variable with parameters n and p (see Definition 427). Then, we have*

$$Q_{1/2} = \begin{cases} \{\lfloor np \rfloor\} & \text{if } p < \frac{1}{2} \\ [\lfloor np \rfloor, \lceil np \rceil] & \text{if } p = \frac{1}{2} \\ \{\lceil np \rceil\} & \text{if } p > \frac{1}{2} \end{cases}, \tag{5.126}$$

where $\lfloor \cdot \rfloor : \mathbb{R} \to \mathbb{Z}$ and $\lceil \cdot \rceil : \mathbb{R} \to \mathbb{Z}$] are the floor and ceiling function, respectively.

Discussion. The distribution function he standard binomial random variable with parameters n and p is given by

$$F_X(x) = \begin{cases} 0 & \text{if } x < x_0 \\ \sum_{j=0}^{k} \binom{n}{j} p^j q^{n-j} & \text{if } x_k \le x < x_{k+1}, \quad \forall k = 0, 1, \ldots . n-1 \\ 1 & \text{if } x_n \le x \end{cases} . \tag{5.127}$$

(see (5.45) of Example 470). Let

$$\hat{k} \equiv \min_{k \in \{0,1,\ldots,n\}} \left\{ \sum_{j=0}^{k} \binom{n}{j} p^j q^{n-j} \ge \frac{1}{2} \right\}.$$

There are two possibilities

$$\sum_{j=0}^{\hat{k}} \binom{n}{j} p^j q^{n-j} = \frac{1}{2} \quad \text{and} \quad \sum_{j=0}^{\hat{k}} \binom{n}{j} p^j q^{n-j} > \frac{1}{2}.$$

In the first case, by virtue of (5.127), we have

$$F_X(x_{\hat{k}}) = \frac{1}{2} \quad \text{and} \quad \lim_{x \to x_{\hat{k}}^-} F_X(x) < \frac{1}{2}.$$

Moreover,

$$F_X(x) = \frac{1}{2} \quad \text{and} \quad \lim_{u \to x^-} F_X(x) = \frac{1}{2},$$

for every $x \in \left(x_{\hat{k}}, x_{\hat{k}+1}\right)$. In the end,

$$F_X\left(x_{\hat{k}+1}\right) > \frac{1}{2} \quad \text{and} \quad \lim_{x \to x_{\hat{k}+1}^-} F_X(x) = \frac{1}{2}.$$

It follows that

$$Q_{1/2} = \left[x_{\hat{k}}, x_{\hat{k}+1}\right].$$

In the second case, still by virtue of (5.127), we have

$$\lim_{x \to x_{\hat{k}}^-} F_X(x) < \frac{1}{2} \quad \text{and} \quad \lim_{u \to x^-} F_X(x) > \frac{1}{2},$$

for every $x \in \left(x_{\hat{k}}, x_{\hat{k}+1}\right)$. As a consequence,

$$Q_{1/2} = \left\{x_{\hat{k}}\right\}.$$

We are left with proving that

$$\begin{array}{ll} x_{\hat{k}} = \lfloor np \rfloor & \text{if } p < \frac{1}{2} \\ \left[x_{\hat{k}}, x_{\hat{k}+1}\right] = \left[\lfloor np \rfloor, \lceil np \rceil\right] & \text{if } p = \frac{1}{2} \\ x_{\hat{k}} = \lceil np \rceil & \text{if } p > \frac{1}{2} \end{array} .$$

This is not an easy task and we do not tackle it here (see Neumann, P., *Über den Median der Binomial and Poissonverteilung*, Wissenschaftliche Zeitschrift der Technischen Universität Dresden (in German), 19: 29–33, (1966), see also Kaas, R.; Buhrman, J.M., *Mean, Median and Mode in Binomial Distributions*, Statistica Neerlandica, 34 (1): 13–18, (1980), doi:10.1111/j.1467-9574.1980.tb00681.x). □

With reference to the standard binomial distribution, note that when $n = 1$ we have

$$\begin{array}{ll} \lfloor np \rfloor = \lfloor p \rfloor = 0 & \text{if } p < \frac{1}{2} \\ \left[\lfloor np \rfloor, \lceil np \rceil \right] = \left[\lfloor p \rfloor, \lceil p \rceil \right] = [0,1] & \text{if } p = \frac{1}{2} \\ \lceil np \rceil = \lceil p \rceil = 1 & \text{if } p > \frac{1}{2} \end{array} .$$

On the other hand,

$$\hat{k} \equiv \min_{k \in \{0,1,\ldots,n\}} \left\{ \sum_{j=0}^{k} \binom{n}{j} p^j q^{n-j} \geq \frac{1}{2} \right\} = \min_{k \in \{0,1\}} \left\{ \sum_{j=0}^{k} p^j q^{1-j} \geq \frac{1}{2} \right\} = \begin{cases} 0 & \text{if } p \leq \frac{1}{2} \\ 1 & \text{if } p > \frac{1}{2} \end{cases} .$$

In addition, if $p = \frac{1}{2}$ we have

$$\sum_{j=0}^{\hat{k}} p^j q^{1-j} = q = \frac{1}{2}.$$

It Then, follows

$$\begin{array}{ll} x_{\hat{k}} = x_0 = 0 & \text{if } p < \frac{1}{2} \\ \left[x_{\hat{k}}, x_{\hat{k}+1} \right] = [x_0, x_1] = [0,1] & \text{if } p = \frac{1}{2} \\ x_{\hat{k}} = x_1 = 1 & \text{if } p > \frac{1}{2} \end{array} .$$

This is the same result as Equation (5.67) of Example 511, which was to be expected in light of the fact that when $n = 1$ the standard binomial distribution becomes the standard Bernoulli distribution.

5.2.7 Mode of a Real Random Variable

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space, let $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu_L) \equiv \mathbb{R}$ be the Euclidean real line equipped with the Borel σ-algebra and the Lebesgue measure μ_L and let $X : \Omega \to \mathbb{R}$ be a real random variable on Ω.

Definition 559 *Assume X is discrete, that is $X(\Omega) \equiv \{x_n\}_{n \in N}$, where $N \subseteq \mathbb{N}$. Then, we call a* mode *of X any $\check{x} \in X(\Omega)$ such that*

$$\check{x} \in \arg\max_{x \in X(\Omega)} \{\mathbf{P}(X = x)\}. \tag{5.128}$$

Definition 560 *Assume X is absolutely continuous with density $f_X : \mathbb{R} \to \mathbb{R}$. Then, we call a* mode *of X any $\check{x} \in \mathbb{R}$ such that*

$$\check{x} \in \arg\operatorname{loc}\max_{x \in \mathbb{R}} \{f_X(x)\}. \tag{5.129}$$

That is any $\check{x} \in \mathbb{R}$ for which there exists $\delta_{\check{x}} > 0$ such that

$$f_X(\check{x}) \geq f_X(x), \quad \forall x \in (\check{x} - \delta_{\check{x}}, \check{x} + \delta_{\check{x}}). \tag{5.130}$$

Definition 561 *The random variable X is said tu be* unimodal *or* multimodal *according to whether X has a unique mode or not.*

Proposition 562 *Assume X is unimodal and symmetric about x_0. Then, x_0 is the mode of X.*

Remark 563 *Assume X is unimodal and symmetric. Then, the mode coincides with the median and the mean, provided that the latter exists.*

5.2.8 Moment of Order One (Expectation) of a Real Random Variable

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space, let $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu_L) \equiv \mathbb{R}$ be the Euclidean real line equipped with the Borel σ-algebra $\mathcal{B}(\mathbb{R})$ and the Lebesgue measure μ_L. Let $X : \Omega \to \mathbb{R}$ be a random variable and let $P_X : \mathbb{R} \to \mathbb{R}$ be the distribution of X.

Definition 564 *We say that X has* finite moment of order one *or* finite expectation *or* finite mean *if the random variable $|X|$ is Lebesgue-integrable on Ω. That is to say,*

$$\int_\Omega |X| \, d\mathbf{P} < \infty.$$

Remark 565 *The random variable X has finite moment of order one if and only if both the positive part X^+ and the negative part X^- of X (see Example 440) have. In this case,*

$$\int_\Omega X \, d\mathbf{P} \overset{def}{=} \int_\Omega X^+ \, d\mathbf{P} - \int_\Omega X^- \, d\mathbf{P}.$$

Proof. The claim follows from the definition of the Lebesgue integral on the probability space Ω. □

Definition 566 *In case X has finite moment of order one, the real number*

$$\mu \equiv \mathbf{E}[X] \overset{def}{=} \int_\Omega X \, d\mathbf{P} \equiv \int_\Omega X(\omega) \, d\mathbf{P}(\omega).$$

is said to be the moment of order one *or the* expectation *or the* expected value *or the* mean *or the* mean value *of X.*

The mean value of a random variable represents the center of the distribution of the random variable in the same sense that the center of mass defined in Physics represents the center of a mass distribution. Indeed, if we think of a probability distribution as a mass distribution with total mass one, Then, the mean value is exactly the center of mass. Not surprisingly, as we will see when introducing the notion of conditional expectation, the mean value of a random variable constitutes the best approximation of the random variable when we cannot observe its realizations in light of the available informetion.

Notation 567 *We write $\mathcal{L}^1(\Omega; \mathbb{R})$ for the set of all real random variables on Ω having finite moment of order one.*

Remark 568 *Let $X \in \mathcal{L}^1(\Omega;\mathbb{R})$ and assume $X \geq 0$. We have*

$$\mathbf{E}[X] \geq 0.$$

Remark 569 *Let $X, Y \in \mathcal{L}^1(\Omega;\mathbb{R})$ and assume $Y \geq X$. We have*

$$\mathbf{E}[Y] \geq \mathbf{E}[Y].$$

Proposition 570 *Let $X \in \mathcal{L}^1(\Omega;\mathbb{R})$. We have*

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

Proposition 571 *For all $X, Y \in \mathcal{L}^1(\Omega;\mathbb{R})$ and all $\alpha, \beta \in \mathbb{R}$ we have $\alpha X + \beta Y \in \mathcal{L}^1(\Omega;\mathbb{R})$ and*

$$\mathbf{E}[\alpha X + \beta Y] = \alpha \mathbf{E}[X] + \beta \mathbf{E}[Y]. \tag{5.131}$$

Proposition 572 *Let X be a real random variable which takes a finite number of values, that is $X(\Omega) \equiv \{x_k\}_{k=1}^n$, for some $n \in \mathbb{N}$ and $x_1, \ldots, x_n \in \mathbb{R}$. (see Proposition **??**). Then, $X \in \mathcal{L}^1(\Omega;\mathbb{R})$ and we have*

$$\mathbf{E}[X] = \sum_{k=1}^n x_k \mathbf{P}(E_k),$$

where $E_k \equiv \{X = x_k\}$, for every $k = 1, \ldots, n$.

Proposition 573 *Let X be an almost surely bounded random variable. Then, $X \in \mathcal{L}^1(\Omega;\mathbb{R})$.*

Example 574 *Given any $x_0 \in \mathbb{R}$, let X be the* Dirac *real random variable concentrated at x_0 (see Definition 406). Then, $X \in \mathcal{L}^1(\Omega;\mathbb{R})$ and we have*

$$\mathbf{E}[X] = x_0.$$

Example 575 *Given any $x_0, x_1 \in \mathbb{R}$, such that $x_0 < x_1$, let X be the Bernoulli random variable with states x_0, x_1 and success probability p (see Definition 416). Then, $X \in \mathcal{L}^1(\Omega;\mathbb{R})$ and we have*

$$\mathbf{E}[X] = x_0 q + x_1 p.$$

In case X is the standard Bernoulli [resp. Rademacher] random variable, we have

$$\mathbf{E}[X] = p \quad [\text{resp. } \mathbf{E}[X] = 0].$$

Proposition 576 *A real random variable $X \in \mathcal{L}^1(\Omega;\mathbb{R})$ if and only if the identity function $id_{\mathbb{R}} : \mathbb{R} \to \mathbb{R}$, given by*

$$id_{\mathbb{R}}(x) \overset{def}{=} x, \quad \forall x \in \mathbb{R},$$

is Lebesgue-integrable on $\mathbb{R}$ with respect to P_X. That is to say,

$$\int_{\mathbb{R}} |x| \, dP_X(x) < \infty. \tag{5.132}$$

In this case, we have

$$\mathbf{E}[X] = \int_{\mathbb{R}} x dP_X(x). \tag{5.133}$$

More generally, if $g : \mathbb{R} \to \mathbb{R}$ is a Borel real function on $\mathbb{R}$, Then, the random variable $g \circ X \in \mathcal{L}^1(\Omega; \mathbb{R})$ if and only if the function g is Lebesgue-integrable on $\mathbb{R}$ with respect to P_X. That is to say,

$$\int_{\mathbb{R}} |g(x)| \, dP_X(x) < \infty. \tag{5.134}$$

In this case, we have

$$\mathbf{E}[g \circ X] = \int_{\mathbb{R}} g(x) \, dP_X(x). \tag{5.135}$$

Proof. ... □

Corollary 577 *Assume the real random variable is absolutely continuous with density $f_X : \mathbb{R} \to \mathbb{R}$. Then, $X \in \mathcal{L}^1(\Omega; \mathbb{R})$ if and only if the product of the functions $id_{\mathbb{R}} : \mathbb{R} \to \mathbb{R}$ and $f_X : \mathbb{R} \to \mathbb{R}$, given by*

$$(id_{\mathbb{R}} \cdot f_X)(x) \overset{def}{=} x f_X(x), \quad \forall x \in \mathbb{R},$$

is Lebesgue-integrable on $\mathbb{R}$ with respect to the Lebesue measure μ_L. That is to say,

$$\int_{\mathbb{R}} |x| \, f_X(x) \, d\mu_L(x) < \infty. \tag{5.136}$$

In this case, we have

$$\mathbf{E}[X] = \int_{\mathbb{R}} x f_X(x) \, d\mu_L(x). \tag{5.137}$$

More generally, if $g : \mathbb{R} \to \mathbb{R}$ is a Borel real function on $\mathbb{R}$, Then, the random variable $g \circ X \in \mathcal{L}^1(\Omega; \mathbb{R})$ if and only if the product of the functions $g : \mathbb{R} \to \mathbb{R}$ and $f_X : \mathbb{R} \to \mathbb{R}$, given by

$$(g \cdot f_X)(x) \overset{def}{=} g(x) f_X(x), \quad \forall x \in \mathbb{R},$$

is Lebesgue-integrable on $\mathbb{R}$ with respect to the Lebesue measure μ_L. That is to say

$$\int_{\mathbb{R}} |g(x)| \, f_X(x) \, d\mu_L(x) < \infty. \tag{5.138}$$

In this case, we have

$$\mathbf{E}[g \circ X] = \int_{\mathbb{R}} g(x) \, f_X(x) \, d\mu_L(x). \tag{5.139}$$

Proof. ... □

Proposition 578 *Let $X \in \mathcal{L}^1(\Omega; \mathbb{R})$ and assume that X is symmetric about $x_0 \in \mathbb{R}$. We have*

$$\mathbf{E}[X] = x_0. \tag{5.140}$$

Proof. By virtue of Proposition 571, both the real random variables $X - x_0$ and $x_0 - X$ have finite moment of order one. In addition, thanks to Equation (5.33), we obtain

$$\mathbf{E}[X - x_0] = \int_{\mathbb{R}} x dP_{X-x_0}(x) = \int_{\mathbb{R}} x dP_{x_0-X}(x) = \mathbf{E}[x_0 - X].$$

From this, applying Equation (5.131), we obtain

$$\mathbf{E}[X] - x_0 = x_0 - \mathbf{E}[X],$$

which cleary implies Equation (5.140). □

5.2.9 Moments of Higher Order, Variance, Skewness, and Kurtosis of a Real Random Variable

The notion of first order moment generalizes immediately to any order $n \geq 1$.

Definition 579 *Given any $n \geq 1$, we say that X has* finite moment of order n *if the random variable $|X|^n$ is Lebesgue-integrable on Ω. That is to say,*

$$\int_{\Omega} |X|^n \, d\mathbf{P} < \infty. \tag{5.141}$$

Definition 580 *In case X has finite moment of order n, the real number*

$$\mathbf{E}[X^n] \equiv \mu'_n \overset{def}{=} \int_{\Omega} X^n \, d\mathbf{P} \equiv \int_{\Omega} X^n(\omega) \, d\mathbf{P}(\omega)$$

*is said to be the (*raw*)* moment of order n *of X.*

We have

$$\begin{array}{ll} \sum_{n \in N} |x_n|^K \mathbf{P}(X = x_n) < \infty & \text{if } X \text{ is discrete} \\ \int_{\mathbb{R}} |x|^K f_X(x) \, d\mu_L(x) < \infty & \text{if } X \text{ is absolutely continuous} \end{array}$$

we set

$$\mu'_K(X) \overset{\text{def}}{=} \begin{cases} \sum_{n \in N} x_n^K \mathbf{P}(X = x_n) & \text{if } X \text{ is discrete} \\ \int_{\mathbb{R}} x^K f_X(x) \, d\mu_L(x) & \text{if } X \text{ is absolutely continuous} \end{cases}$$

Clearly referring to a finite moment of order 1 is the same that referring to a finite moment of order one. Moreover, in terms of notation,

$$\mu'_1 \equiv \mu,$$

where μ is the symbol for the mean or expectation of X.

Notation 581 *We denote by $\mathcal{L}^n(\Omega; \mathbb{R})$ the set of all real random variables on Ω having finite moment of order n.*

Proposition 582 *A real random variable $X \in \mathcal{L}^n(\Omega; \mathbb{R})$ if and only if the nth power function $g : \mathbb{R} \to \mathbb{R}$ given by*

$$g(x) \overset{def}{=} x^n, \quad \forall x \in \mathbb{R},$$

is Lebesgue-integrable on $\mathbb{R}$ with respect to P_X, that is to say

$$\int_{\mathbb{R}} |x^n| \, dP_X(x) < \infty,$$

and in this case we have

$$\mathbf{E}[X^n] = \int_{\mathbb{R}} x^n \, dP_X(x).$$

Example 583 *Given any $x_0 \in \mathbb{R}$, let X be the* Dirac *real random variable concentrated at x_0 (see Definition 406). Then, $X \in \mathcal{L}^n(\Omega; \mathbb{R})$, for every $n \geq 1$, and we have*

$$\mathbf{E}[X^n] = x_0^n.$$

Example 584 *Given any x_0, $x_1 \in \mathbb{R}$, such that $x_0 < x_1$, let X be the Bernoulli random variable with states x_0, x_1 and success probability p (see Definition 416). Then, $X \in \mathcal{L}^n(\Omega;\mathbb{R})$, for every $n \geq 1$, and we have*

$$\mathbf{E}[X^n] = x_0^n q + x_1^n p.$$

In case X is the standard Bernoulli [resp. Rademacher] random variable, we have

$$\mathbf{E}[X^n] = p \quad [\text{resp. } \mathbf{E}[X^n] = \begin{cases} 0 & \text{if } n \text{ odd} \\ 1 & \text{if } n \text{ even} \end{cases}].$$

Remark 585 *Let X be a real random variable which takes a finite number of values $x_1, \ldots, x_m \in \mathbb{R}$ (see Proposition **??**). Then, $X \in \mathcal{L}^n(\Omega;\mathbb{R})$ and we have*

$$\mu_n' = \textstyle\sum_{j=1}^m x_j^n \mathbf{P}(E_j),$$

where $E_j \equiv \{X = x_j\}$, for every $j = 1, \ldots, m$.

Proposition 586 *If $X \in \mathcal{L}^n(\Omega;\mathbb{R})$ Then, $X \in \mathcal{L}^m(\Omega;\mathbb{R})$ for every $m \leq n$.*

Proof. Assume $X \in \mathcal{L}^n(\Omega;\mathbb{R})$ and consider any $m \in \mathbb{N}$ such that $m \leq n$. Since the events $\{|X| < 1\}$ and $\{|X| \geq 1\}$ constitute a decomposition of Ω in mutually incompatible events of $\mathcal{E}$ and the random variable $|X|^m$ is positive, by virtue of the decomposition property of the Lebesgue integral, we can write

$$\int_\Omega |X|^m \, d\mathbf{P} = \int_{\{|X|<1\}} |X|^m \, d\mathbf{P} + \int_{\{|X|\geq 1\}} |X|^m \, d\mathbf{P}. \tag{5.142}$$

Now, on account of the monotonic property of the Lebesgue integral, we have

$$\int_{\{|X|<1\}} |X|^m \, d\mathbf{P} \leq \int_{\{|X|<1\}} 1 \, d\mathbf{P} = \mathbf{P}(|X| < 1) \leq 1 \tag{5.143}$$

and, considering also the monotonic property of the power functions, we can write

$$\int_{\{|X|\geq 1\}} |X|^m \, d\mathbf{P} \leq \int_{\{|X|\geq 1\}} |X|^n \, d\mathbf{P} \leq \int_\Omega |X|^n \, d\mathbf{P} < \infty. \tag{5.144}$$

Combining (5.142)-(5.144), it cleary follows

$$\int_\Omega |X|^m \, d\mathbf{P} < \infty,$$

which means that $X \in \mathcal{L}^m(\Omega;\mathbb{R})$. □

Proposition 587 *If $X \in \mathcal{L}^n(\Omega;\mathbb{R})$ Then, $\alpha X + \beta \in \mathcal{L}^n(\Omega;\mathbb{R})$ for every $\alpha, \beta \in \mathbb{R}$.*

Proof. . □

Corollary 588 *A real random variable $X \in \mathcal{L}^n(\Omega;\mathbb{R})$ if and only if $X - \mathbf{E}[X] \in \mathcal{L}^n(\Omega;\mathbb{R})$.*

Definition 589 *If $X \in \mathcal{L}^n(\Omega;\mathbb{R})$, we call the* central moment of order n *of X the real number*

$$\mu_n \stackrel{def}{=} \mathbf{E}[(X - \mathbf{E}[X])^n] \equiv \mathbf{E}[(X - \mu)^n].$$

Remark 590 *Clearly*

$$\mu_1 = 0.$$

Proposition 591 *A real random variable $X \in \mathcal{L}^n(\Omega;\mathbb{R})$ if and only if the function $g : \mathbb{R} \to \mathbb{R}$, given by*

$$g(x) \overset{def}{=} (x - \mathbf{E}[X])^n, \quad \forall x \in \mathbb{R},$$

is Lebesgue-integrable on $\mathbb{R}$ with respect to P_X and in this case we have

$$\mathbf{E}[(X - \mathbf{E}[X])^n] = \int_{\mathbb{R}} (x - \mathbf{E}[X])^n P_X(dx).$$

Proof. . □

Proposition 592 *Assume $X \in \mathcal{L}^n(\Omega;\mathbb{R})$ for any odd $n > 1$ and that X is symmetric about μ. We have*

$$\mu_n = 0.$$

Proof. The symmetry of X about μ means that $P_{X-\mu} = P_{\mu-X}$ (see (**??**) of Definition 459). This implies

$$\int_{\mathbb{R}} x^n dP_{X-\mu}(x) = \int_{\mathbb{R}} x^n dP_{\mu-X}(x),$$

the two integrals being both finite by virtue of the assumption $X \in \mathcal{L}^n(\Omega;\mathbb{R})$. On the other hand,

$$\int_{\mathbb{R}} x^n dP_{X-\mu}(x) = \int_{\mathbb{R}} (x-\mu)^n dP_X(x) = \mathbf{E}[(X-\mu)^n]$$

and

$$\int_{\mathbb{R}} x^n dP_{\mu-X}(x) = \int_{\mathbb{R}} (\mu-x)^n dP_X(x) = \mathbf{E}[(\mu-X)^n]$$

Since n is odd, it follows

$$\mathbf{E}[(X-\mu)^n] = \mathbf{E}[(\mu-X)^n] = -\mathbf{E}[(X-\mu)^n],$$

which implies our claim. □

Definition 593 *If $X \in \mathcal{L}^2(\Omega;\mathbb{R})$, Then, the central moment of the second order of X is also called the* variance *of X and more usually denoted by the symbols $\mathbf{D}^2[X]$ or $Var(X)$ or σ^2. Hence,*

$$\mathbf{D}^2[X] \equiv Var(X) \equiv \sigma^2 \equiv \mu_2.$$

Proposition 594 *Let $X \in \mathcal{L}^2(\Omega;\mathbb{R})$. We Then, have*

$$\mathbf{D}^2[X] = \mathbf{E}\left[X^2\right] - \mathbf{E}[X]^2.$$

Proof. . □

The variance of a random variables represents the spread of the random variable around its mean. The higher is the variance the more likely the random variable takes its values far from the mean. This fact is the content of the Chebyshev inequality, which will be presented in Subsection 5.3.2.

Example 595 *Given any $x_0 \in \mathbb{R}$, let X be the* Dirac *real random variable concentrated at x_0 (see Definition 406). We have*

$$\mathbf{D}^2[X] = 0.$$

Example 596 *Given any $x_0, x_1 \in \mathbb{R}$, such that $x_0 < x_1$, let X be the Bernoulli random variable with states x_0, x_1 and success probability p (see Definition 416). We have*

$$\mathbf{D}^2[X] = (x_1 - x_0)^2 pq.$$

In case $X : \Omega \to \mathbb{R}$ is the standard Bernoulli [resp. Rademacher] random variable, we have

$$\mathbf{D}^2[X] = pq \quad [\text{resp. } \mathbf{D}^2[X] = 1].$$

Definition 597 *If $X \in \mathcal{L}^2(\Omega; \mathbb{R})$, Then, the arithmetic square root of the variance of X is called the* standard deviation *of X and is usually denoted by the symbols $\mathbf{D}[X]$ or σ. Hence,*

$$\sigma \equiv \mathbf{D}[X] \overset{def}{=} \sqrt{\mathbf{D}^2[X]} \equiv \sqrt{\sigma^2}.$$

Definition 598 *In case X has finite moment of order $n \geq 2$ and $\mathbf{D}^2[X] > 0$, we call the real number*

$$\hat{\mu}_n \overset{def}{=} \frac{\mu_n}{\sigma^n}$$

the standardized central moment of order n *of X.*

Remark 599 *Clearly*

$$\hat{\mu}_2 = 1.$$

Note that the standardized central moments of a random variable are so defined to be invariant under a linear transformation with positive slope, which is a change in the unit of measure of the realizations of the random variable (e.g. degrees Fahrenheit to degrees Celsius). In fact

Proposition 600 *Let $X \in \mathcal{L}^n(\Omega; \mathbb{R})$ and let $\alpha, \beta \in \mathbb{R}$ such that $\beta > 0$. We have*

$$\frac{\mathbf{E}[(\alpha + \beta X - \mathbf{E}[\alpha + \beta X])^n]}{\mathbf{D}[\alpha + \beta X]} = \frac{\mathbf{E}[(X - \mathbf{E}[X])^n]}{\mathbf{D}[X]}.$$

Proof. . □

Standardardized central moments are also useful to obtain an easy comparison between the generating distribution and the standard normal distribution. However, the computation with standardized moments are definitively more difficult.

Definition 601 *If $X \in \mathcal{L}^3(\Omega; \mathbb{R})$ and $\mathbf{D}^2[X] > 0$, Then, we call the central moment [resp. standardized central moment] of the third order of X the* skewness *[resp.* standardized skewness*] of X. The* skewness *[resp.* standardized skewness*] of X is usually denoted by the symbols $Skev(X)$ or γ [resp. $\widehat{Skev}(X)$ or $\hat{\gamma}$]. Hence,*

$$Skev(X) \equiv \gamma \overset{def}{=} \mu_3 \quad [\text{resp. } \widehat{Skev}(X) \equiv \hat{\gamma} \overset{def}{=} \hat{\mu}_3].$$

The real random variable X is said to be positively skewed, negatively skewed *or* unskewed *depending on whether $Skev(X)$ or $\widehat{Skev}(X)$ is positive, negative, or zero.*

Remark 602 *Note that*

$$\widehat{Skev}(X) = \frac{Skev(X)}{\sigma^3}.$$

Example 603 *Given any $x_0, x_1 \in \mathbb{R}$, such that $x_0 < x_1$, let X be the Bernoulli random variable with states x_0, x_1 and success probability p (see Definition 416). We have*

$$Skev(X) = (x_0 - (x_0q + x_1p))^3 q + (x_1 - (x_0q + x_1p))^3 p$$

and

$$\widehat{Skev}(X) = \frac{(x_0 - (x_0q + x_1p))^3 q + (x_1 - (x_0q + x_1p))^3 p}{(x_1 - x_0)^3 p^{\frac{3}{2}} q^{\frac{3}{2}}}.$$

In case $X : \Omega \to \mathbb{R}$ is the standard Bernoulli [resp. Rademacher] random variable, we have

$$Skev(X) = p(1-p)(1-2p) \quad [resp.\ Skev(X) = 0]$$

and

$$\widehat{Skev}(X) = \frac{1-2p}{\sqrt{p(1-p)}} \quad [resp.\ \widehat{Skev}(X) = 0].$$

Loosely speaking, the skewness of a real random variable measures the lack of symmetry of the random variable about the mean value. If the distribution of a real random variable has a longer or ticker tail on the right [resp. on the left] of the mean value, Then, the random variable is positively [resp. negatively] skewed. More precisely, a real random variable which is symmetric about the mean value is unskewed. Nevertheless, an unskewed real random is not necessarily symmetric about the mean value. In fact, we have

Remark 604 *If $X \in \mathcal{L}^3(\Omega; \mathbb{R})$ is symmetric about the mean value, Then, $Skev(X) = 0$.*

Proof. It is sufficient to observe that under the symmetry assumption we have

$$\mu_3 = 0$$

(see Proposition 592). □

Example 605 *The random variable X given by*

$$X = \begin{cases} -\frac{3}{2}, & \mathbf{P}\left(X = -\frac{3}{2}\right) = \frac{312}{2849}, \\ -1, & \mathbf{P}(X = -1) = \frac{289}{1163}, \\ 0, & \mathbf{P}(X = 0) = \frac{39}{128}, \\ \frac{640}{523}, & \mathbf{P}\left(X = \frac{640}{524}\right) = \frac{143055667}{424113536}, \end{cases}$$

is not symmetric and unskewed.

Discussion. By straightforward computations we have

$$\mathbf{E}[X] = 0, \quad \mathbf{E}\left[X^2\right] = 1, \quad \mathbf{E}\left[X^3\right] = 0.$$

Therefore,

$$Skev(X) = \mathbf{E}\left[(X - \mathbf{E}[X])^3\right] = \mathbf{E}\left[X^3\right] = 0.$$

On the other hand, we cleary have

$$P_X \neq P_{-X}.$$

Thus, X is not symmetric about the mean. □

Proposition 606 *Let $X \in \mathcal{L}^3(\Omega;\mathbb{R})$. We have*

$$Skev(X) = \mathbf{E}\left[X^3\right] - 3\mathbf{E}\left[X^2\right]\mathbf{E}[X] + 2\mathbf{E}[X]^3$$

Proof. . □

Definition 607 *If $X \in \mathcal{L}^4(\Omega;\mathbb{R})$ and $\mathbf{D}^2[X] > 0$, we call the central moment [resp. standardized central moment] of the fourth order of X* kurtosis *[resp.* standardized kurtosis*] of X. The kurtosis [resp. standardized kurtosis] of X is usually denoted by the symbols $Kurt(X)$ or κ [resp. $\widehat{Kurt}(X)$ or $\hat{\kappa}$]. Hence,*

$$Kurt(X) \equiv \kappa \overset{def}{=} \mu_4 \quad [resp.\ \widehat{Kurt}(X) \equiv \hat{\kappa} \overset{def}{=} \hat{\mu}_4].$$

The real random variable X is said to be leptokurtic, platykurtic, *or* mesokurtic *depending on whether $\widehat{Kurt}(X) > 3$, $\widehat{Kurt}(X) < 3$, or $\widehat{Kurt}(X) = 3$.*

Example 608 *Given any $x_0, x_1 \in \mathbb{R}$, such that $x_0 < x_1$, let X be the Bernoulli random variable with states x_0, x_1 and success probability p (see Definition 416). We have*

$$Kurt(X) = (x_0 - (x_0q + x_1p))^4 q + (x_1 - (x_0q + x_1p))^4 p.$$

and

$$\widehat{Kurt}(X) = \frac{(x_0 - (x_0q + x_1p))^4 q + (x_1 - (x_0q + x_1p))^4 p}{(x_1 - x_0)^4 p^2 q^2}.$$

In case $X : \Omega \to \mathbb{R}$ is the standard Bernoulli [resp. Rademacher] random variable, we have

$$Kurt(X) = p(1-p)\left(1 - 3p + 3p^2\right) \quad [resp.\ Kurt(X) = \frac{1}{16}]$$

and

$$\widehat{Kurt}(X) = \frac{1 - 3p + 3p^2}{p(1-p)} \quad [resp.\ \widehat{Kurt}(X) = 1].$$

The kurtosis of a real random variable measures the thickness of the tails of the distribution of the random variable: the higher the kurtosis is the more likely the realizations of the extreme values of the random variable are. We will show below that the kurtosis of a standard normal random variable is 3. Therefore, the terms leptokurtic, platykurtic, or mesokurtic account for the thickness of the tail of the distribution of a random variable against a standard normal.

Proposition 609 *Let $X \in \mathcal{L}^4(\Omega;\mathbb{R})$. We have*

$$Kurt(X) = \mathbf{E}\left[X^4\right] - 4\mathbf{E}\left[X^3\right]\mathbf{E}[X] + 6\mathbf{E}\left[X^2\right]\mathbf{E}[X]^2 - 3\mathbf{E}[X]^4.$$

Proof. . □

5.2.10 Discrete Real Random Variables

Among real random variables, *discrete* real random variables play a very important role.

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space and let $X : \Omega \to \mathbb{R}$ be a discrete real $\mathcal{E}$-random variable (see from Definition 420 to Corollary **??**). Setting $X(\Omega) \equiv \{x_n\}_{n \in N}$, where $N \subseteq \mathbb{N}$, and $E_n \equiv \{X = x_n\}$, we know that $E_n \in \mathcal{E}$, for every $n \in N$. Furthermore, considering the distribution $P_X : \mathcal{B}(\mathbb{R}) \to \mathbb{R}_+$ of X, we can write

$$P_X(x_n) \equiv \mathbf{P}(X = x_n) \equiv \mathbf{P}(E_n),$$

for every $\forall n \in N$. As a consequence,

Proposition 610 *We have*

$$P_X(B) \equiv \mathbf{P}(X \in B) = \sum_{\{n \in N : x_n \in B\}} P_X(x_n) \equiv \sum_{\{n \in N : x_n \in B\}} \mathbf{P}(X = x_n) \equiv \sum_{\{n \in N : x_n \in B\}} \mathbf{P}(E_n),$$

equivalently

$$P_X(B) = \sum_{n \in N} P_X(x_n) 1_B(x_n) \equiv \sum_{n \in N} \mathbf{P}(X = x_n) 1_B(x_n) \equiv \sum_{n \in N} \mathbf{P}(E_n) 1_B(x_n),$$

*for every $B \in \mathcal{B}(\mathbb{R})$ (see Example **??** and **??**).*

Example 611 *Let $\Omega \equiv \{\omega_{1,1}, \omega_{1,2}, \omega_{2,1}, \ldots, \omega_{6,6}\}$ be the sample space of all possible outcomes of the roll of two fair dice, and let $X : \Omega \to \mathbb{R}$ be the function given by*

$$X(\omega_{j,k}) \stackrel{def}{=} j + k, \quad \forall j, k = 1, \ldots, 6.$$

We have $X(\Omega) = \{2, 3, \ldots, 12\}$ and setting

$$E_n \equiv \{X = x_n\}, \quad \forall n = 2, \ldots, 12$$

we can clearly write

$$X = \sum_{n=2}^{12} n 1_{E_n}.$$

Therefore, given a σ-algebra of events $\mathcal{E}$ on Ω, the function X is a real $\mathcal{E}$–random variable if and only if $E_n \in \mathcal{E}$, for every $n = 2, \ldots, 12$. In this case, with reference to the naive probability $\mathbf{P} : \mathcal{E} \to \mathbb{R}_+$, we have

$$p_n \equiv \mathbf{P}(X = n) \equiv \mathbf{P}(E_n) = \begin{cases} \frac{1}{11}, & \text{if } n = 2, 12, \\ \frac{2}{11}, & \text{if } n = 3, 11, \\ \frac{3}{11}, & \text{if } n = 4, 10, \\ \frac{4}{11}, & \text{if } n = 5, 9, \\ \frac{5}{11}, & \text{if } n = 6, 8, \\ \frac{6}{11}, & \text{if } n = 7. \end{cases}$$

Discussion. Note that

$$\begin{aligned} E_2 &\equiv \{X = 2\} = \{\omega_{1,1}\}, \\ E_3 &\equiv \{X = 3\} = \{\omega_{1,2}, \omega_{2,1}\}, \\ E_4 &\equiv \{X = 4\} = \{\omega_{1,3}, \omega_{2,2}, \omega_{3,1}\}, \\ &\dots \\ E_{12} &\equiv \{X = 12\} = \{\omega_{6,6}\}. \end{aligned}$$

Now, since the dice are fair, with reference to the naive probability $\mathbf{P} : \mathcal{E} \to \mathbb{R}_+$, we can write

$$\mathbf{P}(E_n) = \frac{|E_n|}{|\Omega|}$$

for every $n = 2, \dots, 12$. We clearly have $|\Omega| = 11$. In addition, from the conditions $1 \leq j, k \leq 6$, $j + k = n$, it easily follows

$$|E_n| = |\{j : ((k-6) \vee 1) \leq j \leq ((k-1) \wedge 6)\}|,$$

The latter implies

$$|E_n| = \begin{cases} 1, & \text{if } k = 2, 12, \\ 2, & \text{if } k = 3, 11, \\ 3, & \text{if } k = 4, 10, \\ 4, & \text{if } k = 5, 9, \\ 5, & \text{if } k = 6, 8, \\ 6, & \text{if } k = 7, \end{cases}$$

for every $n = 2, \dots, 12$, which yields the desired distribution. □

Proposition 612 *A a point $x_{1/2} \in \mathbb{R}$ a median for X if and only if we have*

$$\sum\nolimits_{\{n \in N \,:\, x_n \leq x_{1/2}\}} P_X(x_n) \geq \frac{1}{2} \quad \text{and} \quad \sum\nolimits_{\{n \in N \,:\, x_n \geq x_{1/2}\}} P_X(x_n) \geq \frac{1}{2}.$$

Proof. . □

In terms of expectation and variance of discrete real random variables we can state

Proposition 613 *For any $n \in \mathbb{N}$, we have $X \in \mathcal{L}^n(\Omega; \mathbb{R})$ if and only if*

$$\sum_{k \in N} |x_k|^n \mathbf{P}(E_k) < \infty. \tag{5.145}$$

In this case, we have

$$\mathbf{E}[X^n] = \sum_{k \in K} x_k^n \mathbf{P}(E_k). \tag{5.146}$$

and

$$\mathbf{E}[(X - \mathbf{E}[X])^n] = \sum_{k \in K} (x_k - \mathbf{E}[X])^n \mathbf{P}(E_k). \tag{5.147}$$

Proof. The family $\{E_k\}_{k\in K}$ is a partition of Ω, moreover with no loss of generality we can assume that $\mathbf{P}(E_k) > 0$ for every $k \in K$. Hence, by virtue of the properties of the Lebesgue integral, we have

$$\int_\Omega |X|^n \, d\mathbf{P} = \int_{\bigcup_{k\in K} E_k} |X|^n \, d\mathbf{P} = \sum_{k\in K} \int_{E_k} |X|^n \, d\mathbf{P} = \sum_{k\in K} \int_{E_k} |x_k|^n \, d\mathbf{P} = \sum_{k\in K} |x_k|^n \, \mathbf{P}(E_k).$$

Therefore, the necessity and sufficiency of (5.145) immediately follows. Now, under (5.145), on account of Proposition 586, we have

$$\int_\Omega |X|^m \, d\mathbf{P} = \sum_{k\in K} |x_k|^m \, \mathbf{P}(E_k) < \infty$$

for every $m \leq n$. It follows

$$\begin{aligned}\int_\Omega |X - \mathbf{E}[X]|^n \, d\mathbf{P} &= \int_\Omega \left| \textstyle\sum_{m=0}^n (-1)^m \binom{n}{m} \mathbf{E}[X]^m X^{n-m} \right| d\mathbf{P} \\ &\leq \int_\Omega \textstyle\sum_{m=0}^n \binom{n}{m} |\mathbf{E}[X]|^m |X|^{n-m} \, d\mathbf{P} \\ &= \textstyle\sum_{m=0}^n \binom{n}{m} |\mathbf{E}[X]|^m \int_\Omega |X|^{n-m} \, d\mathbf{P} < \infty.\end{aligned}$$

As a consequence, we can write

$$\begin{aligned}\mathbf{E}[(X - \mathbf{E}[X])^n] &= \int_\Omega (X - \mathbf{E}[X])^n \, d\mathbf{P} = \int_{\bigcup_{k\in K} E_k} (X - \mathbf{E}[X])^n \, d\mathbf{P} \\ &= \sum_{k\in K} \int_{E_k} (X - \mathbf{E}[X])^n \, d\mathbf{P} = \sum_{k\in K} \int_{E_k} (x_k - \mathbf{E}[X])^n \, d\mathbf{P} = \\ &= \sum_{k\in K} (x_k - \mathbf{E}[X])^n \, \mathbf{P}(E_k),\end{aligned}$$

which proves (5.147). □

Corollary 614 *For any $n \in \mathbb{N}$, we have $X \in \mathcal{L}^n(\Omega;\mathbb{R})$ if and only if*

$$\sum_{k\in K} |x_k|^n \, P_X(x_k) < \infty. \tag{5.148}$$

In this case, we have

$$\mathbf{E}[X^n] = \sum_{k\in K} x_k^n P_X(x_k) \tag{5.149}$$

and

$$\mathbf{E}[(X - \mathbf{E}[X])^n] = \sum_{k\in K} (x_k - \mathbf{E}[X])^n \, P_X(x_k). \tag{5.150}$$

More generally, if $g : \mathbb{R} \to \mathbb{R}$ is a Borel-measurable real function on $\mathbb{R}$, Then, the random variable $g \circ X \in \mathcal{L}^n(\Omega;\mathbb{R})$ if and only if

$$\sum_{k\in K} |g(x_k)|^n \, P_X(x_k) < \infty. \tag{5.151}$$

In this case, we have

$$\mathbf{E}\left[(g\circ X)^n\right]=\sum_{n\geq 0}g(x_k)^nP_X\left(x_k\right). \tag{5.152}$$

Proof. . □

Binomial random variable with success probability p

Let $(\Omega,\mathcal{E},\mathbf{P})\equiv\Omega$ be a probability space and, for some $n\in\mathbb{N}$, let $(E_k)_{k=0}^n$ be a partition of Ω.

Example 615 *Assume*

$$\mathbf{P}\left(E_k\right)\equiv\binom{n}{k}p^kq^{n-k},\quad k=0,1,\ldots,n, \tag{5.153}$$

where $0<p<1$ and $q\equiv 1-p$. Fix any strictly increasing finite sequence $(x_k)_{k=0}^n$ of real number and consider the function $X:\Omega\to\mathbb{R}$ given by

$$X\left(\omega\right)\overset{def}{=}\sum_{k=0}^n x_k1_{E_k}\left(\omega\right),\quad\forall\omega\in\Omega. \tag{5.154}$$

The function $X:\Omega\to\mathbb{R}$ is a discrete real $\mathcal{E}$-random variable. The distribution $P_X:\mathcal{B}\left(\mathbb{R}\right)\to\mathbb{R}_+$ of X is given by

$$P_X(B)=\sum_{k=0}^n\binom{n}{k}p^kq^{n-k}1_B\left(x_k\right),\quad\forall B\in\mathcal{B}\left(\mathbb{R}\right). \tag{5.155}$$

Discussion. The function X is a discrete real $\mathcal{E}$-random variable by virtue of Proposition (**??**). In addition, according to the definition of X, we have

$$\begin{aligned}P_X(B)&=\mathbf{P}\left(X\in B\right)=\mathbf{P}\left(\left(X\in B\right)\cap\left(\bigcup_{k=0}^nE_k\right)\right)=\mathbf{P}\left(\bigcup_{k=0}^n\left(\left(X\in B\right)\cap E_k\right)\right)\\&=\sum_{k=0}^n\mathbf{P}\left(\left(X\in B\right)\cap E_k\right)=\sum_{k=0}^n\mathbf{P}\left(E_k\right)1_B\left(x_k\right).\end{aligned} \tag{5.156}$$

In fact,

$$\left(X\in B\right)\cap E_k=\begin{cases}E_k & \text{if } x_k\in B\\ \varnothing & \text{if } x_k\in B\end{cases}.$$

Hence,

$$\mathbf{P}\left(\left(X\in B\right)\cap E_k\right)=\mathbf{P}\left(E_k\right)1_B\left(x_k\right)$$

(see also Proposition **??**). Combining (5.153) with (5.156), the desired (5.155) immediately follows. □

Definition 616 *We call the random variable introduced in Example **??** the* binomial *random variable with states $(x_k)_{k=0}^n$ and success probability $p\equiv\mathbf{P}\left(X=x_1\right)$. In case $x_k\equiv k$, for every $k=0,1,\ldots,n$, we speak of the* standard binomial *random variable.*

Proposition 617 *Let $X : \Omega \to \mathbb{R}$ a binomial random variable with states $(x_k)_{k=0}^k$ and success probability p. We have*

$$\mathbf{E}[X] = \sum_{k=0}^{n} x_k \binom{n}{k} p^k q^{n-k}, \tag{5.157}$$

and

$$\mathbf{D}^2[X] = \sum_{k=0}^{n} (x_k - \mathbf{E}[X])^2 \binom{n}{k} p^k q^{n-k}. \tag{5.158}$$

In particular, for a standard binomial random variable we have

$$\mathbf{E}[X] = np \tag{5.159}$$

and

$$\mathbf{D}^2[X] = npq. \tag{5.160}$$

Proof. Thanks to Proposition 572, **??** and Corollary 614, we need to compute only,

$$\sum_{k=0}^{n} x_k P_X(x_k) = \sum_{k=0}^{n} x_k \binom{n}{k} p^k q^{n-k},$$

and

$$\sum_{k=0}^{n} (x_k - \mathbf{E}[X])^2 P_X(x_k) = \sum_{k=0}^{n} (x_k - \mathbf{E}[X])^2 \binom{n}{k} p^k q^{n-k}.$$

Now, in the particular case, we have

$$\begin{aligned}
\sum_{k=0}^{n} k \binom{n}{k} p^k q^{n-k} &= \sum_{k=1}^{n} k \binom{n}{k} p^k q^{n-k} \\
&= \sum_{k=1}^{n} k \frac{n!}{k!(n-k)!} p^k q^{n-k} \\
&= \sum_{k=1}^{n} \frac{n!}{(k-1)!(n-k)!} p^k q^{n-k} \\
&= np \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!((n-1)-(k-1))!} p^{k-1} q^{(n-1)-(k-1)} \\
&= np \sum_{\ell=0}^{n-1} \frac{(n-1)!}{\ell!((n-1)-\ell)!} p^j q^{(n-1)-\ell} \\
&= np \sum_{\ell=0}^{n-1} \binom{n-1}{\ell} p^j q^{(n-1)-\ell} \\
&= np(p+q)^{n-1} \\
&= np,
\end{aligned}$$

which proves (5.159). Finally,

$$\sum_{k=0}^{n} (k - np)^2 \binom{n}{k} p^k q^{n-k} = \sum_{k=0}^{n} k^2 \binom{n}{k} p^k q^{n-k} - \sum_{k=0}^{n} 2knp \binom{n}{k} p^k q^{n-k} + \sum_{k=0}^{n} n^2 p^2 \binom{n}{k} p^k q^{n-k},$$

where

$$\sum_{k=0}^{n} 2knp\binom{n}{k}p^k q^{n-k} = 2np\sum_{k=0}^{n} k\binom{n}{k}p^k q^{n-k} = 2n^2p^2, \tag{5.161}$$

and

$$\sum_{k=0}^{n} n^2p^2\binom{n}{k}p^k q^{n-k} = n^2p^2\sum_{k=0}^{n}\binom{n}{k}p^k q^{n-k} = n^2p^2. \tag{5.162}$$

In addition,

$$\begin{aligned}
\sum_{k=0}^{n} k^2\binom{n}{k}p^k q^{n-k} &= \sum_{k=1}^{n} k^2\binom{n}{k}p^k q^{n-k} \\
&= \sum_{k=1}^{n} k\frac{n!}{(k-1)!(n-k)!}p^k q^{n-k} \\
&= \sum_{k=1}^{n}((k-1)+1)\frac{n!}{(k-1)!(n-k)!}p^k q^{n-k} \\
&= \sum_{k=1}^{n}(k-1)\frac{n!}{(k-1)!(n-k)!}p^k q^{n-k} + \sum_{k=1}^{n}\frac{n!}{(k-1)!(n-k)!}p^k q^{n-k},
\end{aligned}$$

where

$$\sum_{k=1}^{n}\frac{n!}{(k-1)!(n-k)!}p^k q^{n-k} = np$$

and

$$\begin{aligned}
\sum_{k=1}^{n}(k-1)\frac{n!}{(k-1)!(n-k)!}p^k q^{n-k} &= \sum_{k=2}^{n}(k-1)\frac{n!}{(k-1)!(n-k)!}p^k q^{n-k} \\
&= \sum_{k=2}^{n}\frac{n!}{(k-2)!(n-k)!}p^k q^{n-k} \\
&= n(n-1)p^2\sum_{k=2}^{n}\frac{(n-2)!}{(k-2)!((n-2)-(k-2))!}p^{k-2}q^{(n-2)-(k-2)} \\
&= n(n-1)p^2\sum_{\ell=0}^{n-2}\frac{(n-2)!}{\ell!((n-2)-\ell)!}p^j q^{(n-2)-\ell} \\
&= n(n-1)p^2\sum_{\ell=0}^{n-2}\binom{n-2}{\ell}p^k q^{(n-2)-\ell} \\
&= n(n-1)p^2(p+q)^{n-2} \\
&= n(n-1)p^2.
\end{aligned}$$

Therefore,

$$\sum_{k=0}^{n} k^2\binom{n}{k}p^k q^{n-k} = n(n-1)p^2 + np. \tag{5.163}$$

Thus, combining (5.161), (5.162), and (5.163), the desired (5.160) easily follows[4]. □

[4]In what follows, we will show a much simpler way to obtain both (5.159) and (5.160) by means of the notion of independent random variables.

Example 618 *Consider an urn containing $N \geq 2$ balls of which M are white, with $1 \leq M \leq N$, and $N - M$ are black. A sample of $n \geq 1$ balls is drawn from the urn* with replacement *of the drawn ball in the urn. Denote by X the random variable counting the number of white balls in the sample. We have*

$$\mathbf{P}(X = m) = \binom{n}{m} p^m q^{n-m}, \quad \forall m = 0, 1, \ldots, n,$$

where

$$p \equiv \frac{M}{N}, \qquad q \equiv 1 - p = \frac{N - M}{N}.$$

Hence, X has a standard binomial distribution with success probability p.

Bivariate hypergeometric random variable with success probability p

Let $M, N \in \mathbb{N}$ such that $M \leq N$, let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space and, for some $n \leq N$, let $(E_m)_{m=0}^n$ be a partition of Ω.

Example 619 *Assume*

$$\mathbf{P}(E_m) \overset{def}{=} \begin{cases} \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}} & \text{if } m \leq M \\ 0 & \text{if } m > M \end{cases}. \tag{5.164}$$

Fix any strictly increasing finite sequence $(x_m)_{m=0}^n$ of real number and consider the function $X : \Omega \to \mathbb{R}$ given by

$$X(\omega) \overset{def}{=} \sum_{m=0}^{n} x_m 1_{E_m}(\omega), \quad \forall \omega \in \Omega. \tag{5.165}$$

The function $X : \Omega \to \mathbb{R}$ is a discrete real $\mathcal{E}$-random variable. The distribution $P_X : \mathcal{B}(\mathbb{R}) \to \mathbb{R}_+$ of X is given by

$$P_X(B) = \sum_{m=0}^{n \wedge M} \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}} 1_B(x_m), \quad \forall B \in \mathcal{B}(\mathbb{R}). \tag{5.166}$$

Discussion. . □

Definition 620 *We call the random variable introduced in Example 619 the* bivariate hypergeometric *random variable with states $(x_m)_{m=0}^n$ and success probability $p \equiv \frac{M}{N}$. In case $x_m \equiv m$, for every $m = 0, 1, \ldots, n$, we speak of the* standard bivariate hypergeometric *random variable.*

Proposition 621 *Let $X : \Omega \to \mathbb{R}$ a bivariate hypergeometric random variable with states $(x_m)_{m=0}^n$ and success probability p. We have*

$$\mathbf{E}[X] = \sum_{m=0}^{n \wedge M} x_m \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}}, \tag{5.167}$$

and

$$\mathbf{D}^2[X] = \sum_{m=0}^{n \wedge M} (x_m - \mathbf{E}[X])^2 \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}}. \tag{5.168}$$

In particular, for a standard bivariate hypergeometric random variable we have

$$\mathbf{E}[X] = np \tag{5.169}$$

and

$$\mathbf{D}^2[X] = np\frac{(N-M)(N-n)}{N(N-1)}. \tag{5.170}$$

Proof. By virtue of the identities

$$m\binom{M}{m} = M\binom{M-1}{m-1}, \quad m \geq 1 \qquad \text{and} \qquad \binom{N}{n} = \frac{N}{n}\binom{N-1}{n-1}, \quad N \geq n \geq 1, \tag{5.171}$$

we obtain

$$\begin{aligned}
\mathbf{E}[X] &= \textstyle\sum_{m=0}^{n\wedge M} m\frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}} = \sum_{m=1}^{n\wedge M}\frac{M\binom{M-1}{m-1}\binom{N-1-(M-1)}{n-1-(m-1)}}{\frac{N}{n}\binom{N-1}{n-1}} \\
&= \frac{nM}{N}\textstyle\sum_{m=1}^{n\wedge M}\frac{\binom{M-1}{m-1}\binom{N-1-(M-1)}{n-1-(m-1)}}{\binom{N-1}{n-1}} = np\sum_{j=0}^{n-1\wedge M-1}\frac{\binom{M-1}{j}\binom{N-1-(M-1)}{n-1-j}}{\binom{N-1}{n-1}} \\
&= np,
\end{aligned}$$

because the Vandermonde identity implies

$$\textstyle\sum_{j=0}^{n-1\wedge M-1}\frac{\binom{M-1}{j}\binom{N-1-(M-1)}{n-1-j}}{\binom{N-1}{n-1}} = 1. \tag{5.172}$$

Now, still on account of (5.171) and (5.172), we can write

$$\begin{aligned}
\mathbf{E}\left[X^2\right] &= \textstyle\sum_{m=0}^{n\wedge M} m^2\frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}} = \sum_{m=1}^{n\wedge M} m\frac{M\binom{M-1}{m-1}\binom{N-1-(M-1)}{n-1-(m-1)}}{\frac{N}{n}\binom{N-1}{n-1}} \\
&= \frac{nM}{N}\textstyle\sum_{j=0}^{n-1\wedge M-1}(j+1)\frac{\binom{M-1}{j}\binom{N-1-(M-1)}{n-1-j}}{\binom{N-1}{n-1}} \\
&= np\left(\textstyle\sum_{j=0}^{n-1\wedge M-1} j\frac{\binom{M-1}{j}\binom{N-1-(M-1)}{n-1-j}}{\binom{N-1}{n-1}} + \sum_{j=0}^{n-1\wedge M-1}\frac{\binom{M-1}{j}\binom{N-1-(M-1)}{n-1-j}}{\binom{N-1}{n-1}}\right) \\
&= np\left(\textstyle\sum_{j=1}^{n-1\wedge M-1}\frac{(M-1)\binom{M-2}{j-1}\binom{N-2-(M-2)}{n-2-(j-1)}}{\frac{N-1}{n-1}\binom{N-2}{n-2}} + 1\right) \\
&= np\left(\frac{(M-1)(n-1)}{N-1}\textstyle\sum_{h=0}^{n-2\wedge M-2}\frac{\binom{M-2}{h}\binom{N-2-(M-2)}{n-2-h}}{\binom{N-2}{n-2}} + 1\right) \\
&= np\left(\frac{(M-1)(n-1)}{N-1} + 1\right).
\end{aligned}$$

As a consequence,

$$\begin{aligned}
\mathbf{D}^2[X] &= np\left(\frac{(M-1)(n-1)}{N-1} + 1\right) - n^2p^2 \\
&= np\left(\frac{(M-1)(n-1)}{N-1} + 1 - \frac{nM}{N}\right) \\
&= np\frac{(N-M)(N-n)}{N(N-1)}
\end{aligned}$$

as desired. □

Example 622 *Consider an urn containing $N \geq 2$ balls of which M are white, with $1 \leq M \leq N$, and $N - M$ are black. A sample of $n \leq N$ balls is drawn from the urn* without replacement *of the drawn ball in the urn. Denote by X the random variable counting the number of white balls in the sample. We have*

$$\mathbf{P}(X = m) = \begin{cases} \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}} & \text{if } m \leq M \\ 0 & \text{if } m > M \end{cases}.$$

Hence, X has a standard bivariate hypergeometric distribution with success probability $p \equiv \frac{M}{N}$.

Geometric random variable with success probability p

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space and let $(E_n)_{n\geq 1}$ be a partition of Ω.

Example 623 *Assume*

$$\mathbf{P}(E_n) \overset{def}{=} pq^{n-1}, \quad \forall n \geq 1, \tag{5.173}$$

where $0 < p < 1$, $q \equiv 1 - p$. Fix any strictly increasing finite sequence $(x_n)_{n\geq 1}$ of real number and consider the function $X : \Omega \to \mathbb{R}$ given by

$$X(\omega) \overset{def}{=} \sum_{n=1}^{\infty} x_n 1_{E_n}(\omega), \quad \forall \omega \in \Omega. \tag{5.174}$$

The function $X : \Omega \to \mathbb{R}$ is a discrete real $\mathcal{E}$-random variable. The distribution $P_X : \mathcal{B}(\mathbb{R}) \to \mathbb{R}_+$ of X is given by

$$P_X(B) = \sum_{n=1}^{\infty} pq^{n-1} 1_B(x_n), \quad \forall B \in \mathcal{B}(\mathbb{R}). \tag{5.175}$$

Discussion. . □

Definition 624 *We call the random variable introduced in Example 623 the* geometric *random variable with states $(x_n)_{n\geq 1}$ and success probability p. In case $x_n \equiv n$, for every $n \geq 1$, we speak of the* standard geometric *random variable.*

Proposition 625 *Let $X : \Omega \to \mathbb{R}$ be a geometric random variable with states $(x_n)_{n\geq 1}$ and success probability p. Then, X has a finite moment of order one [resp. two] if and only if*

$$\sum_{n=1}^{\infty} |x_n| pq^{n-1} < \infty \quad \text{[resp. } \sum_{n=1}^{\infty} x_n^2 pq^{n-1} < \infty\text{]}.$$

In this case, we have

$$\mathbf{E}[X] = \sum_{n=1}^{\infty} x_n pq^{n-1} \quad \text{[resp. } \mathbf{D}^2[X] = \sum_{n=1}^{\infty} (x_n - \mathbf{E}[X])^2 pq^{n-1}\text{]}.$$

In particular, for a standard geometric random variable we have

$$\mathbf{E}[X] = \frac{1}{p}$$

and

$$\mathbf{D}^2[X] = \frac{1-p}{p^2}.$$

Proof. We have

$$\begin{aligned}\mathbf{E}[X] &= \sum_{n=1}^{\infty} npq^{n-1} = p\sum_{n=1}^{\infty} nq^{n-1} = p\sum_{n=1}^{\infty}\frac{d}{dq}q^n = p\frac{d}{dq}\sum_{n=1}^{\infty} q^n \\ &= p\frac{d}{dq}\left(q\sum_{n=1}^{\infty} q^{n-1}\right) = p\frac{d}{dq}\left(q\sum_{m=0}^{\infty} q^m\right) = p\frac{d}{dq}\left(\frac{q}{1-q}\right) \\ &= p\frac{1-q+q}{(1-q)^2} = \frac{1}{p}.\end{aligned}$$

Similarly,

$$\begin{aligned}\mathbf{E}\left[X^2\right] &= \sum_{n=1}^{\infty} n^2 pq^{n-1} = p\sum_{n=1}^{\infty} n^2 q^{n-1} = p\sum_{n=1}^{\infty}\frac{d}{dq}\left(nq^n\right) = p\frac{d}{dq}\sum_{n=1}^{\infty} nq^n \\ &= p\frac{d}{dq}\left(q\sum_{n=1}^{\infty} nq^{n-1}\right) = p\frac{d}{dq}\left(q\sum_{n=1}^{\infty}\frac{d}{dq}q^n\right) = p\frac{d}{dq}\left(q\frac{d}{dq}\sum_{n=1}^{\infty} q^n\right) \\ &= p\frac{d}{dq}\left(q\frac{d}{dq}\frac{q}{1-q}\right) = p\frac{d}{dq}\left(\frac{q}{(1-q)^2}\right) \\ &= p\left(\frac{(1-q)^2 + 2q(1-q)}{(1-q)^4}\right) = p\left(\frac{1-q+2q}{(1-q)^3}\right) \\ &= p\left(\frac{p+2(1-p)}{p^3}\right) = \frac{2-p}{p^2}\end{aligned}$$

As a consequence,

$$\mathbf{D}^2[X] = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}$$

as desired. □

Example 626 *Consider an urn containing $N \geq 2$ balls of which M are white, with $1 \leq M \leq N$, and $N - M$ are black. A sample of balls is drawn from the urn* with replacement *of the drawn ball in the urn until a white ball is drawn. Denote by X the random variable counting the draw in which the white ball appears. We have*

$$\mathbf{P}(X = n) = pq^{n-1}, \quad \forall n \geq 1,$$

where $0 < p < 1$, $q \equiv 1 - p$. Hence, X has a standard geometric distribution with success probability p.

Poisson random variable with intensity parameter λ

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space and let $(E_n)_{n\geq 1}$ be a partition of Ω.

Example 627 *Assume*

$$\mathbf{P}(E_n) \equiv e^{-\lambda}\frac{\lambda^n}{n!}, \quad n \geq 0, \tag{5.176}$$

where $\lambda > 0$. Fix any strictly increasing finite sequence $(x_n)_{n\geq 0}$ of real number and consider the function $X : \Omega \to \mathbb{R}$ given by

$$X(\omega) \overset{def}{=} \sum_{n=0}^{\infty} x_n 1_{E_n}(\omega), \quad \forall \omega \in \Omega. \tag{5.177}$$

The function $X : \Omega \to \mathbb{R}$ is a discrete real $\mathcal{E}$-random variable. The distribution $P_X : \mathcal{B}(\mathbb{R}) \to \mathbb{R}_+$ of X is given by

$$P_X(B) = \sum_{n=0}^{\infty} e^{-\lambda} \frac{\lambda^n}{n!} 1_B(x_n), \quad \forall B \in \mathcal{B}(\mathbb{R}). \tag{5.178}$$

Discussion. . □

Definition 628 *We call the random variable introduced in Example 627 the* Poisson *random variable with states $(x_n)_{n\geq 0}$ and rate or intensity parameter λ. In case $x_n \equiv n$, for every $n \geq 0$, we speak of the* standard Poisson *random variable.*

Notation 629 *It is customary to denote the standard Poisson random variable wth rate parameter λ by the symbol $Poiss(\lambda)$.*

5.2.11 Absolutely Continuous Real Random Variables

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space, let $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu_L) \equiv \mathbb{R}$ be the Euclidean real line equipped with the Borel σ-algebra $\mathcal{B}(\mathbb{R})$ and the Lebesgue measure μ_L, and let $X : \Omega \to \mathbb{R}$ be a real $\mathcal{E}$-random variable on Ω with distribution $P_X : \mathcal{B}(\mathbb{R}) \to \mathbb{R}_+$.

Recall 630 *The random variable X is said to be absolutely continuous if the distribution P_X is absolutely continuous with respecet to μ_L. In this case the random variable X has a density $f_X : \mathbb{R} \to \mathbb{R}$ such that*

$$P_X(B) = \int_B f_X(x)\, d\mu_L(x).$$

The following result that should be compared with Proposition (**??**).

Proposition 631 *An absolutely continuous random variable $X \in \mathcal{L}^1(\Omega; \mathbb{R})$ if and only if the Borel-measurable function $h : \mathbb{R} \to \mathbb{R}$ given by*

$$h(x) \overset{def}{=} x f_X(x), \quad \forall x \in \mathbb{R},$$

is Lebesgue-integrable on $\mathbb{R}$, that is to say

$$\int_{\mathbb{R}} |x f_X(x)|\, d\mu_L(x) < \infty,$$

and in this case we have

$$\mathbf{E}[X] = \int_{\mathbb{R}} x f_X(x)\, d\mu_L(x).$$

More generally, if $g : \mathbb{R} \to \mathbb{R}$ is a Borel-measurable real function on $\mathbb{R}$, Then, the random variable $g \circ X \in \mathcal{L}^1(\Omega; \mathbb{R})$ if and only if the Borel-measurable function $h : \mathbb{R} \to \mathbb{R}$ given by

$$h(x) \overset{def}{=} g(x) f_X(x), \quad \forall x \in \mathbb{R},$$

is Lebesgue integrable on $\mathbb{R}$*, that is to say*

$$\int_{\mathbb{R}} |g(x) f_X(x)| \, d\mu_L(x) < \infty,$$

and in this case we have

$$\mathbf{E}[g \circ X] = \int_{\mathbb{R}} g(x) f_X(x) \, d\mu_L(x).$$

As a special case, $X \in \mathcal{L}^n(\Omega; \mathbb{R})$ *if and only if the Borel-measurable function* $h : \mathbb{R} \to \mathbb{R}$*, given by*

$$h(x) \stackrel{def}{=} x^n f_X(x), \quad \forall x \in \mathbb{R},$$

is Lebesgue-integrable on $\mathbb{R}$ *and in this case we have*

$$\mathbf{E}[X^n] = \int_{\mathbb{R}} x^n f_X(x) \, d\mu_L(x).$$

Proof. . □

Corollary 632 *If* X *is absolutely continuous and* $X \in \mathcal{L}^n(\Omega; \mathbb{R})$*, Then, the Borel-measurable function* $h : \mathbb{R} \to \mathbb{R}$ *given by*

$$h(x) \stackrel{def}{=} (x - \mathbf{E}[X])^n f_X(x), \quad \forall x \in \mathbb{R},$$

is Lebesgue-integrable on $\mathbb{R}$ *and in this case we have*

$$\mu_n = \int_{\mathbb{R}} (x - \mathbf{E}[X])^n f_X(x) \, d\mu_L(x).$$

Proof. . □

Let $X : \Omega \to \mathbb{R}$ be an absolutely continuous real $\mathcal{E}$-random variable on Ω with density $f_X : \mathbb{R} \to \mathbb{R}$.

Uniform density

Definition 633 *Fixed any* $a, b \in \mathbb{R}$ *with* $a < b$*, we say that* X *is* continuous uniformly distributed *in the interval* $[a, b]$*, and we write* $X \sim Unif(a, b)$*, if* f_X *is the continuous uniform density in the interval* $[a, b]$ *given by Equation (4.73). That is*

$$f_X(x) \stackrel{def}{=} \frac{1}{b-a} 1_{[a,b]}(x), \quad \forall x \in \mathbb{R}. \tag{5.179}$$

Proposition 634 *We have*

$$F_X(x) = \frac{x-a}{b-a} 1_{[a,b]}(x) + 1_{(b,+\infty)}(x), \quad \forall x \in \mathbb{R}. \tag{5.180}$$

Proof. In fact, considering (5.179), we can write

$$F_X(x) = \int_{(-\infty,x]} f_X(u)\, d\mu_L(u) = \int_{(-\infty,x]} \frac{1}{b-a} 1_{[a,b]}(x) d\mu_L(u) = \frac{1}{b-a} \int_{(-\infty,x]\cap[a,b]} d\mu_L(u)$$
$$= \begin{cases} \frac{1}{b-a} \int_{\varnothing} d\mu_L(u) & \text{if } x < a \\ \frac{1}{b-a} \int_{[a,x]} d\mu_L(u) & \text{if } a \leq x \leq b \\ \frac{1}{b-a} \int_{[a,b]} d\mu_L(u) & \text{if } b < x \end{cases}.$$

On the other hand,

$$\int_B d\mu_L(u) = \mu_L(B),$$

for every $B \in \mathcal{B}(\mathbb{R})$. Hence,

$$F_X(x) = \begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b-a} & \text{if } a \leq x \leq b \\ 1 & \text{if } b < x \end{cases},$$

which is the desired result. □

Proposition 635 *We have*

$$\mathbf{E}[X] = \frac{1}{2}(a+b) \qquad \textit{and} \qquad \mathbf{D}^2[X] = \frac{1}{12}(a-b)^2.$$

Proof. Staightforward computations yield

$$\begin{aligned} \mathbf{E}[X] &= \int_{\mathbb{R}} x \cdot \frac{1}{b-a} 1_{[a,b]}(x)\, d\mu_L(x) \\ &= \frac{1}{b-a} \int_{[a,b]} x\, d\mu_L(x) = \frac{1}{b-a} \int_a^b x\, dx \\ &= \frac{1}{b-a} \frac{1}{2}(b^2 - a^2) = \frac{1}{2}(b+a). \end{aligned}$$

and

$$\begin{aligned} \mathbf{E}\left[X^2\right] &= \int_{\mathbb{R}} x^2 \cdot \frac{1}{b-a} 1_{[a,b]}(x)\, d\mu_L(x) \\ &= \frac{1}{b-a} \int_{[a,b]} x^2\, d\mu_L(x) = \frac{1}{b-a} \int_a^b x^2\, dx \\ &= \frac{1}{3} \frac{b^3 - a^3}{b-a} = \frac{1}{3}\left(a^2 + ab + b^2\right). \end{aligned}$$

As a consequence,

$$\mathbf{D}^2[X] = \frac{1}{3}\left(a^2 + ab + b^2\right) - \frac{1}{4}(a+b)^2 = \frac{1}{12}(a-b)^2.$$

□

Cauchy Density

Definition 636 *We say that X is* Cauchy distributed *if f_X is the uniform density introduced in Definition 332. That is to say*

$$f_X(x) \overset{def}{=} \frac{1}{\pi}\frac{1}{1+x^2}, \quad \forall x \in \mathbb{R}.$$

Proposition 637 *A Cauchy distributed random variable has no moments of any order.*

Proof. A straightforward computation gives

$$\begin{aligned}\int_\Omega |X|\, d\mathbf{P} &= \int_{\mathbb{R}} \frac{1}{\pi}\frac{|x|}{1+x^2} d\mu_L(x) = \frac{2}{\pi}\int_{\mathbb{R}_+} \frac{x}{1+x^2} d\mu_L(x) \\ &= \frac{2}{\pi}\lim_{x\to+\infty}\int_0^x \frac{u}{1+u^2}du = \frac{1}{\pi}\lim_{x\to+\infty}\int_0^x \frac{1}{1+u^2} d\left(1+u^2\right) \\ &= \frac{1}{\pi}\lim_{x\to+\infty}\int_1^{1+x^2}\frac{1}{v}dv = \frac{1}{\pi}\lim_{x\to+\infty}\ln(v)\big|_1^{1+x^2} \\ &= \frac{1}{\pi}\lim_{x\to+\infty}\ln\left(1+x^2\right) = +\infty.\end{aligned}$$

This, on account of Proposition 586, proves the claim. □

Exponential Density

Definition 638 *Given $\lambda > 0$, we say that X is* exponentially distributed *with parameter λ, and we write $X \sim Exp(\lambda)$, if f_X is the exponential density introduced in Definition 335. That is*

$$f_X(x) \overset{def}{=} \lambda e^{-\lambda x} 1_{\mathbb{R}_+}(x), \quad \forall x \in \mathbb{R}.$$

Proposition 639 *An exponentially distributed random variable with parameter λ has finite moments of any order given by*

$$\mathbf{E}[X^n] = \frac{n!}{\lambda^n}, \quad \forall n \geq 1.$$

In particular, we have

$$\mathbf{E}[X] = \frac{1}{\lambda} \qquad \text{and} \qquad \mathbf{D}^2[X] = \frac{1}{\lambda^2}.$$

Proof. For any $n \geq 1$, we have

$$\int_\Omega |X|^n\, d\mathbf{P} = \int_{\mathbb{R}} |x|^n \lambda e^{-\lambda x} 1_{\mathbb{R}_+}(x) d\mu_L(x) = \lambda \int_{\mathbb{R}_+} x^n e^{-\lambda x} d\mu_L(x) = \lambda \lim_{x\to+\infty}\int_0^x u^n e^{-\lambda u} du. \tag{5.181}$$

On the other hand, integrating by parts

$$\int u^n e^{-\lambda u} du = -\frac{1}{\lambda}\int u^n de^{-\lambda u} = -\frac{1}{\lambda}\left(u^n e^{-\lambda u} - \int e^{-\lambda u} du^n\right) = -\frac{1}{\lambda}\left(u^n e^{-\lambda u} - n\int u^{n-1} e^{-\lambda u} du\right).$$

Hence,

$$\lim_{x\to+\infty}\int_0^x u^n e^{-\lambda u}du = -\frac{1}{\lambda}\lim_{x\to+\infty}\left(u^n e^{-\lambda u}\Big|_0^x - n\int_0^x u^{n-1}e^{-\lambda u}du\right) = \frac{n}{\lambda}\lim_{x\to+\infty}\int_0^x u^{n-1}e^{-\lambda u}du.$$

Iterating the procedure, it is not difficult to show that

$$\lim_{x\to+\infty}\int_0^x u^n e^{-\lambda u}du = \frac{n!}{\lambda^n}\lim_{x\to+\infty}\int_0^x e^{-\lambda u}du, \tag{5.182}$$

where

$$\lim_{x\to+\infty}\int_0^x e^{-\lambda u}du = \frac{1}{\lambda}. \tag{5.183}$$

This implies that X has finite moment of any order $n \geq 1$. In addition, combining (5.181)-(5.183), the above computation yields

$$\mathbf{E}\left[X^n\right] = \int_\Omega X^n d\mathbf{P} = \int_\Omega |X|^n \, d\mathbf{P} = \frac{n!}{\lambda^n},$$

for any $n \geq 1$. In particular,

$$\mathbf{E}\left[X\right] = \frac{1}{\lambda}$$

and

$$\mathbf{D}^2\left[X\right] = \mathbf{E}\left[X^2\right] - \mathbf{E}\left[X\right]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

as desired. □

Laplace Density

Definition 640 *Given $m, \sigma \in \mathbb{R}$ with $\sigma > 0$, we say that X is* Lapace distributed *with parameters μ and σ, and we write $X \sim Laplace\,(\mu, \sigma)$, if f_X is the exponential density introduced in Definition 338. That is*

$$f_X(x) \overset{def}{=} \frac{1}{2\sigma}e^{-\frac{|x-\mu|}{\sigma}}, \quad \forall x \in \mathbb{R}.$$

Proposition 641 *We have*

$$\mathbf{E}\left[X\right] = \mu \qquad \textit{and} \qquad \mathbf{D}^2\left[X\right] = 2\sigma^2.$$

Proof. . □

Normal Density

Definition 642 *Given $\mu, \sigma \in \mathbb{R}$ with $\sigma > 0$, we say that X is* normally *or* Gaussian distributed *with parameters μ and σ, and we write $X \sim N(\mu, \sigma)$, if f_X is the exponential density introduced in Definition* **??**. *That is*

$$f_X(x) \overset{def}{=} \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad \forall x \in \mathbb{R}.$$

In particular, when $\mu = 0$ and $\sigma = 1$, we refer to $X \sim N(0,1)$ as a standard normal *random variable.*

Proposition 643 *We have*

$$\mathbf{E}[X] = \mu \tag{5.184}$$

and

$$\mathbf{D}^2[X] = \sigma^2. \tag{5.185}$$

Proof. Setting $y \equiv (x-\mu)/\sqrt{2}\sigma$, we can write

$$\begin{aligned}
\mathbf{E}[X] &= \int_{-\infty}^{+\infty} x f_X(x)\, dx \\
&= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{+\infty} \frac{x}{\sqrt{2}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx \\
&= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{+\infty} (\sqrt{2}\sigma y + \mu) e^{-y^2}\, dy \\
&= \frac{1}{\sqrt{\pi}} \lim_{x\to+\infty} \int_{-x}^{x} (\sqrt{2}\sigma y + \mu) e^{-y^2}\, dy \\
&= \frac{1}{\sqrt{\pi}} \lim_{x\to+\infty} \left(\sqrt{2}\sigma \int_{-x}^{x} y e^{-y^2}\, dy + \mu \int_{-x}^{x} e^{-y^2}\, dy \right) \\
&= \frac{1}{\sqrt{\pi}} \mu \int_{-\infty}^{\infty} e^{-y^2}\, dy \\
&= \mu.
\end{aligned}$$

Similarly, setting again $y \equiv (x-\mu)/\sqrt{2}\sigma$ we can write

$$\begin{aligned}
\mathbf{E}\left[X^2\right] &= \int_{-\infty}^{+\infty} x^2 f_X(x)\, dx \\
&= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2}\sigma} x^2 e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx \\
&= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{+\infty} (\sqrt{2}\sigma y + \mu)^2 e^{-y^2}\, dy \\
&= \frac{1}{\sqrt{\pi}} \lim_{x\to+\infty} \int_{-x}^{x} (\sqrt{2}\sigma y + \mu)^2 e^{-y^2}\, dy \\
&= \frac{1}{\sqrt{\pi}} \lim_{x\to+\infty} \left(2\sigma^2 \int_{-x}^{x} y^2 e^{-y^2}\, dy + 2\sqrt{2}\sigma \int_{-x}^{x} y e^{-y^2}\, dy + \mu^2 \int_{-x}^{x} e^{-y^2}\, dy \right).
\end{aligned}$$

On the other hand, integrating by parts,

$$\int_{-x}^{x} y^2 e^{-y^2}\, dy = \frac{1}{2}\left(-y e^{-y^2}|_{-t}^{t} + \int_{-x}^{x} e^{-y^2}\, dy \right) = \frac{1}{2}\int_{-x}^{x} e^{-y^2}\, dy.$$

It Then, follows

$$\int_{-\infty}^{+\infty} x^2 f_X(x)\, dx = \frac{1}{\sqrt{\pi}}\left(\sigma^2 + \mu^2\right) \lim_{x\to+\infty} \int_{-x}^{x} e^{-y^2}\, dy = \sigma^2 + \mu^2.$$

As a consequence,

$$\mathbf{D}^2[X] = \mathbf{E}\left[X^2\right] - \mathbf{E}[X]^2 = \int_{-\infty}^{+\infty} x^2 f_X(x)\, dx - \left(\int_{-\infty}^{+\infty} x f_X(x)\, dx \right)^2 = \sigma^2.$$

This completes the proof. □

Definition 644 *The distribution function of a standard normal random variable is also commonly known as the* cumulative distribution function *(CDF) of the standard normal density and it is denoted by the special symbol* Φ*. We can Then, write*

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{y^2}{2}} dy, \quad \forall x \in \mathbb{R}.$$

Remark 645 *We have*

$$\Phi(-x) = 1 - \Phi(x), \quad \forall x \in \mathbb{R}. \tag{5.186}$$

In particular,

$$\Phi(0) = 1/2.$$

Proof. For any $x \in \mathbb{R}$, by the change of variable $y \equiv -z$, we can write

$$\Phi(-x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-x} e^{-\frac{y^2}{2}} dy = -\frac{1}{\sqrt{2\pi}} \int_{+\infty}^{x} e^{-\frac{z^2}{2}} dz = \frac{1}{\sqrt{2\pi}} \int_{x}^{+\infty} e^{-\frac{z^2}{2}} dz.$$

As a consequence,

$$\Phi(x) + \Phi(-x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{y^2}{2}} dy + \frac{1}{\sqrt{2\pi}} \int_{x}^{+\infty} e^{-\frac{z^2}{2}} dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{y^2}{2}} dy = 1.$$

The desired result immediately follows.

Remark 646 *Assume* $X \sim N(0,1)$*. Then*

$$-X \sim N(0,1). \tag{5.187}$$

Proof. For any $x \in \mathbb{R}$, we have

$$\mathbf{P}(-X \leq x) = \mathbf{P}(X \geq -x) = 1 - \mathbf{P}(X < -x) = 1 - \mathbf{P}(X \leq -x) = 1 - \Phi(-x).$$

Thanks to (5.186), it Then, follows

$$\mathbf{P}(-X \leq x) = \Phi(x),$$

which yields the desired result. □

Remark 647 *Assume* $X \sim N(\mu, \sigma^2)$*. We have*

$$\int_{-\infty}^{+\infty} \left| x^n e^{-\frac{(x-\mu)^2}{2\sigma^2}} \right| dx < \infty, \quad \forall n \geq 0.$$

Therefore, a $N(\mu, \sigma^2)$ *random variable has finite moments of all orders.*

Proof. It is clearly sufficient to prove Remark 647 for n even, that is $n = 2k$, for $k = n/2$. In this case, we can write

$$\int_{-\infty}^{+\infty} \left| x^n e^{-\frac{(x-\mu)^2}{2\sigma^2}} \right| dx = \int_{-\infty}^{+\infty} \left| x^{2k} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \right| dx = \int_{-\infty}^{+\infty} x^{2k} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \lim_{t \to +\infty} \int_{-t}^{t} x^{2k} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx.$$

On the other hand, since the density $f_{\mu,\sigma}(x)$ is quickly decreasing as x goes to infinite, we have

$$\lim_{x\to+\infty}\frac{x^{2k}e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{1/x^2}=\lim_{x\to+\infty}x^{2k+2}e^{-\frac{(x-\mu)^2}{2\sigma^2}}=\sqrt{2\pi}\sigma\lim_{x\to+\infty}x^{2k+2}f_{\mu,\sigma}(x)=0.$$

It Then, follows that

$$\lim_{a\to+\infty}\int_{-a}^{+a}x^{2k}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx<\infty,$$

as desired.

Remark 648 *Assume $\mu\equiv 0$, Then, we have*

$$\mathbf{E}\left[X^{2k+1}\right]=0,\quad\forall k\geq 0, \tag{5.188}$$

and

$$\mathbf{E}\left[X^{2k}\right]=1\cdot 3\cdot\cdots\cdot(2k-1)\,\sigma^{2k},\quad\forall k\geq 0. \tag{5.189}$$

Proof. We have

$$(-x)^{2k+1}\exp\left(-\frac{(-x)^2}{2\sigma^2}\right)=-x^{2k+1}\exp\left(-\frac{x^2}{2\sigma^2}\right),\quad\forall k\geq 0.$$

This clearly proves (5.188). Now, for $k=0$ we have

$$\mathbf{E}\left[X^{2k}\right]=\mathbf{E}\left[X^0\right]=\mathbf{E}\left[1\right]=1.$$

In addition, considering the inductive assumption that (5.189) holds true for some $k>0$, integrating by parts, we have

$$\begin{aligned}\mathbf{E}\left[X^{2k+2}\right]&=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}x^{2k+2}e^{-\frac{x^2}{2\sigma^2}}\,dx\\&=-\sigma^2\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}x^{2k+1}\,d\left(e^{-\frac{x^2}{2\sigma^2}}\right)\\&=-\sigma^2\frac{1}{\sigma\sqrt{2\pi}}\left[\lim_{x\to+\infty}y^{2k+1}e^{-\frac{y^2}{2\sigma^2}}\Big|_{-x}^{x}-(2k+1)\int_{-\infty}^{\infty}x^{2k}e^{-\frac{x^2}{2\sigma^2}}\,dx\right]\\&=(2k+1)\,\sigma^2\mathbf{E}\left[X^{2k}\right]\\&=1\cdot 3\cdot\cdots\cdot(2k-1)\,(2k+1)\,\sigma^{2k+2}.\end{aligned}$$

Therefore, by the induction principle we obtain (5.189), as desired. □

Remark 649 *Assume $X\sim N(0,1)$ and let $\mu,\sigma\in\mathbb{R}$ with $\sigma>0$. Then, we have*

$$\mu+\sigma X\sim N\left(\mu,\sigma^2\right)$$

Proof. For any $x\in\mathbb{R}$, we have

$$\mathbf{P}\left(\mu+\sigma X\leq x\right)=\mathbf{P}\left(X\leq(x-\mu)/\sigma\right)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\frac{x-\mu}{\sigma}}e^{-\frac{y^2}{2}}dy.$$

On the other hand, by the change of variable $y \equiv (z-\mu)\,\sigma$, we can write

$$\int_{-\infty}^{\frac{x-\mu}{\sigma}} e^{-\frac{y^2}{2}} dy = \frac{1}{\sigma}\int_{-\infty}^{x} e^{-\frac{(z-\mu)^2}{2\sigma^2}} dz.$$

It Then, follows

$$\mathbf{P}\left(\mu + \sigma X \leq x\right) = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x} e^{-\frac{(z-\mu)^2}{2\sigma^2}} dz,$$

which guarantees that the random variables $\mu + \sigma X$ and $N\left(\mu, \sigma^2\right)$ have the same distribution. □

5.3 Inequalities for Real Random Variables

5.3.1 Markov Inequalities

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space, let $(\mathbb{R}, \mathcal{B}(\mathbb{R})) \equiv \mathbb{R}$ be the real Borel state space, and let $X : \Omega \to \mathbb{R}$ be a real random variable.

Theorem 650 (I Markov inequality) *Assume*

1. $X \geq 0$,
2. X *has finite expectation and* $[X] > 0$.

Then, we have

$$\mathbf{P}(X \geq \lambda \mathbf{E}\left[X\right]) \leq \frac{1}{\lambda}, \tag{5.190}$$

for every $\lambda > 0$.

Proof. Fixed any $\lambda > 0$, the events

$$\{X < \lambda \mathbf{E}\left[X\right]\} \quad \text{and} \quad \{X \geq \lambda \mathbf{E}\left[X\right]\}$$

constitute a partition of Ω. Hence, thanks to the additive and monotone properties of the Lebesgue integral and the positivity of X, we can write

$$\begin{aligned}\mathbf{E}\left[X\right] &= \int_{\Omega} X d\mathbf{P} = \int_{\{X<\lambda\mathbf{E}[X]\}} X d\mathbf{P} + \int_{\{X\geq\lambda\mathbf{E}[X]\}} X d\mathbf{P} \geq \int_{\{X\geq\lambda\mathbf{E}[X]\}} X d\mathbf{P} \\ &\geq \int_{\{X\geq\lambda\mathbf{E}[X]\}} \lambda\mathbf{E}\left[X\right] d\mathbf{P} = \lambda\mathbf{E}\left[X\right]\mathbf{P}\left(X \geq \lambda\mathbf{E}\left[X\right]\right),\end{aligned}$$

Dividing the first and last member of the above inequality chain by $\lambda\mathbf{E}\left[X\right]$, the desired result follows. □

Corollary 651 *Under the same assumptions of Theorem 650, we have*

$$\mathbf{P}(X \geq \kappa) \leq \frac{\mathbf{E}\left[X\right]}{\kappa}, \tag{5.191}$$

for every $\kappa > 0$.

Proof. Fixed any $\kappa > 0$, referring Equation (5.190) to $\lambda \equiv \kappa / \mathbf{E}[X]$, we can write

$$\mathbf{P}(X \geq \kappa) = \mathbf{P}\left(X \geq \frac{\kappa}{\mathbf{E}[X]}\mathbf{E}[X]\right) \leq \frac{1}{\kappa/\mathbf{E}[X]},$$

which is the desired result. □

Example 652 *Assume X satisfies the assumptions of Theorem 650. Assume further that the probability distribution P_X of X is unknown. Given $\kappa > 0$,we want to determine an upper bound for $\mathbf{P}(X > \kappa)$.*

Discussion. Note that if P_X were known, the distribution function F_X of X would be known as well and we could write

$$\mathbf{P}(X > \kappa) = 1 - \mathbf{P}(X \leq \kappa) = 1 - P_X((-\infty, \kappa]) = 1 - F_X(x),$$

which would be a sharp evaluation of $\mathbf{P}(X > \kappa)$. However, when the probability distribution P_X of X is unknown, we are in a position to apply Corollary 651 and we can write

$$\mathbf{P}(X > \kappa) \leq \mathbf{P}(X \geq \kappa) \leq \frac{\mathbf{E}[X]}{\kappa},$$

which yields a possible upper bound for $\mathbf{P}(X > \kappa)$. □

Example 653 *Assume to have an urn containing some balls of different weights. Assume to know that the average weight of the balls in the urn is $2g$. We want to determine an upper bound for the probability that the weight of a ball, randomly drawn from the urn, is not smaller than $3g$.*

Discussion. Write Ω for the sample space representing the urn, the sample points $\omega \in \Omega$ being the single balls. We can represent the outcomes of the draws by means of the values taken a random variable $X : \Omega \to \mathbb{R}$ having for states the weights of the balls in the urn, that is

$$X(\omega) \stackrel{\text{def}}{=} weight(\omega), \quad \forall \omega \in \Omega.$$

The assumptions of Corollary 651 are satisfied. Therefore, we can write

$$\mathbf{P}(X \geq 3) \leq \frac{\mathbf{E}[X]}{3} = \frac{2}{3},$$

which is the desired result. □

Theorem 654 (II Markov inequality) *With no additional assumptions on the random variable $X : \Omega \to \mathbb{R}$, consider a non-decreasing Borel positive function $\varphi : \mathbb{R} \to \bar{\mathbb{R}}_+$ such that the random variable $\varphi \circ X : \Omega \to \bar{\mathbb{R}}_+$ has finite expectation. Then*

$$\varphi(x)\,\mathbf{P}(X \geq x) \leq \mathbf{E}[\varphi(X)], \tag{5.192}$$

for every $x \in \mathbb{R}$.

Proof. For any $x \in \mathbb{R}$, the events

$$\{X < x\} \quad \text{and} \quad \{X \geq x\}$$

constitute a partition of Ω. Thanks to the additive property of the Lebesgue integral and the positivity of φ, we can then write

$$\mathbf{E}\left[\varphi\left(X\right)\right] = \int_{\Omega} \varphi\left(X\right) d\mathbf{P} = \int_{\{X<x\}} \varphi\left(X\right) d\mathbf{P} + \int_{\{X\geq x\}} \varphi\left(X\right) d\mathbf{P} \geq \int_{\{X\geq x\}} \varphi\left(X\right) d\mathbf{P}, \quad (5.193)$$

Now, since φ is non-decreasing, the monotone property of the Lebesgue integral implies

$$\int_{\{X\geq x\}} \varphi\left(X\right) d\mathbf{P} \geq \int_{\{X\geq x\}} \varphi\left(x\right) d\mathbf{P} = \varphi\left(x\right) \mathbf{P}(X \geq x). \quad (5.194)$$

Combining (5.193) and (5.194), we obtain

$$\mathbf{E}\left[\varphi\left(X\right)\right] \geq \varphi\left(x\right) \mathbf{P}(X \geq x),$$

which is the desired (5.192). □

5.3.2 Chebyshev Inequality

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space, let $(\mathbb{R}, \mathcal{B}(\mathbb{R})) \equiv \mathbb{R}$ be the real Borel state space, and let X be a real random variable on Ω.

Theorem 655 *Assume $\varphi : \mathbb{R} \to \mathbb{R}$ is a Borel positive function of real variable such that*

1. *$\varphi\left(x\right) = \varphi\left(-x\right)$ for every $x \in \mathbb{R}$;*
2. *$\varphi\left(x\right) \leq \varphi\left(y\right)$ for all $0 \leq x \leq y$;*
3. *the random variable $\varphi \circ X : \Omega \to \bar{\mathbb{R}}_{+}$ has finite expectation.*

Then, we have

$$\varphi\left(x\right) \mathbf{P}\left(\left|X\right| \geq x\right) \leq \mathbf{E}\left[\varphi\left(X\right)\right], \quad (5.195)$$

for every $x \geq 0$.

Proof. For any fixed $x \geq 0$, the events $\{|X| < x\}$ and $\{|X| \geq x\}$ constitute a partition of Ω. Hence, thanks to the additive property of the Lebesgue integral and the positivity of φ, we can write

$$\mathbf{E}\left[\varphi\left(X\right)\right] \overset{\text{def}}{=} \int_{\Omega} \varphi\left(X\right) d\mathbf{P} = \int_{\{|X|<x\}} \varphi\left(X\right) d\mathbf{P} + \int_{\{|X|\geq x\}} \varphi\left(X\right) d\mathbf{P} \geq \int_{\{|X|\geq x\}} \varphi\left(X\right) d\mathbf{P}. \quad (5.196)$$

On the other hand, we have

$$\{|X| \geq x\} = \{X \leq -x\} \cup \{X \geq x\}.$$

This, on account of 1 2, allows to show that for every $\omega \in \{|X| \geq x\}$ we have $\varphi\left(X\left(\omega\right)\right) \geq \varphi\left(x\right)$. In fact, in the case $\omega \in \{X \geq x\}$, we have $X\left(\omega\right) \geq x$ and, since $x \geq 0$, from 2, it follows

$\varphi(X(\omega)) \geq \varphi(x)$. In the case $\omega \in \{X \leq -x\}$, we have $X(\omega) \leq -x$, which implies $-X(\omega) \geq x$. Again, from 2, it follows $\varphi(-X(\omega)) \geq \varphi(x)$ and, applying 1, we end up with $\varphi(X(\omega)) \geq \varphi(x)$. As a consequence, thanks to the monotone property of the Lebesgue integral, we obtain

$$\int_{\{|X|\geq x\}} \varphi(X)\, d\mathbf{P} \geq \int_{\{|X|\geq x\}} \varphi(x)\, d\mathbf{P} = \varphi(x)\, \mathbf{P}(|X| \geq x), \tag{5.197}$$

Combining (5.196) and (5.197), the desired (5.195) follows. □

Corollary 656 *Assume $X : \Omega \to \mathbb{R}$ has finite moment of order 2 and $\mathbf{E}[X^2] > 0$. Then, we have*

$$\mathbf{P}(|X - \mathbf{E}[X]| \geq \kappa) \leq \frac{\mathbf{D}^2[X]}{\kappa^2}, \quad \forall \kappa > 0. \tag{5.198}$$

Proof. Corollary 656 follows either from Corollary 651 or Theorem 655. In fact, applying Corollary 651 to the random variable $Y \equiv (X - \mathbf{E}[X])^2$, we can write

$$\mathbf{P}\left((X - \mathbf{E}[X])^2 \geq \kappa^2\right) \leq \frac{\mathbf{E}\left[(X - \mathbf{E}[X])^2\right]}{\kappa^2} = \frac{\mathbf{D}^2[X]}{\kappa^2}, \tag{5.199}$$

for any $\kappa > 0$. On the other hand, observing that

$$(X - \mathbf{E}[X])^2 \geq \kappa^2 \Leftrightarrow |X - \mathbf{E}[X]| \geq \kappa,$$

we have

$$\mathbf{P}\left((X - \mathbf{E}[X])^2 \geq \kappa^2\right) = \mathbf{P}(|X - \mathbf{E}[X]| \geq \kappa). \tag{5.200}$$

Thus, combining equations (5.199) and (5.200), we obtain the desired 5.198. Similarly, applying Theorem 655 to the random variable $Y = X - \mathbf{E}[X]$ and the function $\varphi : \mathbb{R} \to \mathbb{R}$ given by

$$\varphi(x) \stackrel{\text{def}}{=} x^2,$$

we can write

$$\mathbf{E}\left[(X - \mathbf{E}[X])^2\right] \geq \kappa^2 \mathbf{P}(|X - \mathbf{E}[X]| \geq \kappa)$$

for any $\kappa \geq 0$, and again the desired 5.198 follows. □

Example 657 *Assume $X : \Omega \to \mathbb{R}$ has finite moment of order 2. Assume further that the probability distribution $P_X : \mathcal{B}(\mathbb{R}) \to \mathbb{R}_+$ of X is unknown. Then, fixed any $a, b \in \mathbb{R}$ such that $a < \mathbf{E}[X] < b$, we can write*

$$\mathbf{P}(a < X < b) > 1 - \frac{\mathbf{D}[X^2]}{\kappa^2}, \tag{5.201}$$

where $\kappa \equiv \min\{\mathbf{E}[X] - a, b - \mathbf{E}[X]\}$.

Discussion. Note that if the distribution $P_X : \mathcal{B}(\mathbb{R}) \to \mathbb{R}_+$ of X were known, the distribution function $F_X : \mathbb{R} \to \mathbb{R}$ of X would be known as well and we could write

$$\mathbf{P}(a < X < b) = P_X(a, b) = \lim_{x \to b^-} F_X(x) - F_X(a).$$

This would be a sharp evaluation of $\mathbf{P}(a < X < b)$. However, if we ignore $F_X : \mathbb{R} \to \mathbb{R}$, under the assumptions considered, we are in a position to apply Corollary 656. We have

$$a < X < b \Leftrightarrow a - \mathbf{E}[X] < X - \mathbf{E}[X] < b - \mathbf{E}[X]. \tag{5.202}$$

Furthermore, setting $\kappa \equiv \min\{\mathbf{E}[X] - a, b - \mathbf{E}[X]\}$, we have

$$[-\kappa, \kappa] \subseteq [a - \mathbf{E}[X], b - \mathbf{E}[X]].$$

Hence,

$$|X - \mathbf{E}[X]| < \kappa \Leftrightarrow -\kappa < X - \mathbf{E}[X] < \kappa \Rightarrow a - \mathbf{E}[X] < X - \mathbf{E}[X] < b - \mathbf{E}[X]. \quad (5.203)$$

Combining (5.202) and (5.202), it follows

$$\{|X - \mathbf{E}[X]| < \kappa\} \subseteq \{a < X < b\}.$$

This implies

$$\mathbf{P}(a < X < b) \geq \mathbf{P}(|X - \mathbf{E}[X]| < \kappa). \quad (5.204)$$

On the other hand,

$$\mathbf{P}(|X - \mathbf{E}[X]| < \kappa) = 1 - \mathbf{P}(|X - \mathbf{E}[X]| \geq \kappa)$$

and applying Equation (5.198), we obtain

$$\mathbf{P}(|X - \mathbf{E}[X]| < \kappa) \geq 1 - \frac{\mathbf{D}[X^2]}{\kappa^2}. \quad (5.205)$$

In the end, combining (5.204) and (5.205), the desired (5.201) follows. □

5.3.3 Jensen Inequality

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space, let $(\mathbb{R}, \mathcal{B}(\mathbb{R})) \equiv \mathbb{R}$ be the real Borel state space, let $X : \Omega \to \mathbb{R}$ be a real random variable, and let $\varphi : \mathbb{R} \to \mathbb{R}$ be a Borel convex function of real variable.

Theorem 658 (Jensen Inequality) *Assume the random variables$X$ and $\varphi \circ X : \Omega \to \mathbb{R}$ have finite expectation. Then, we have*

$$\varphi(\mathbf{E}[X]) \leq \mathbf{E}[\varphi(X)] \quad (5.206)$$

Proof. Right now, we are in a position to prove only that the Jensen inequality holds true for simple random variable. We will be able to complete the proof for all the random variables satisfying the assumption of Theorem 658, once we introduce the notion of convergence for sequences of random variables and related properties. Therefore, assume for now that X is a simple random variable. Then, we can write

$$X = \sum_{k=1}^{n} x_k 1_{E_k},$$

where $\{x_1, \ldots, x_n\} \equiv X(\Omega)$ and $E_k \equiv \{X = x_k\}$, for every $k = 1, \ldots, n$. We have

$$\sum_{k=1}^{n} \mathbf{P}(E_k) = 1 \quad \text{and} \quad \mathbf{E}[X] = \sum_{k=1}^{n} x_k \mathbf{P}(E_k).$$

Hence, thanks to the convexity of ϕ, we can write

$$\phi\left(\mathbf{E}\left[X\right]\right)=\phi\left(\sum_{k=1}^{n}x_{k}\mathbf{P}\left(E_{k}\right)\right)\leq\sum_{k=1}^{n}\phi\left(x_{k}\right)\mathbf{P}\left(E_{k}\right)=\mathbf{E}\left[\phi\left(X\right)\right]. \tag{5.207}$$

This proves Equation (5.206) for simple randon variables. In what follows, we will see thay for any random variable $X:\Omega\rightarrow\mathbb{R}$ having finite expectation there always exists an increasing sequence $\left(X_{n}\right)_{n\geq1}$ of simple random variables *converging to* X *almost surely* on Ω and satisfying $\left|X_{n}\right|\leq\left|X\right|$ for every $n\geq1$. For such a sequence, we will prove that

$$\lim_{n\rightarrow\infty}\mathbf{E}\left[X_{n}\right]=\mathbf{E}\left[X\right]$$

and from the convexity of ϕ, which implies that ϕ is locally Lipschitz continuous, will follow

$$\lim_{n\rightarrow\infty}\phi\left(\mathbf{E}\left[X_{n}\right]\right)=\phi\left(\mathbf{E}\left[X\right]\right). \tag{5.208}$$

Now, by virtue of (5.207), we will have

$$\phi\left(\mathbf{E}\left[X_{n}\right]\right)\leq\mathbf{E}\left[\phi\left(X_{n}\right)\right], \tag{5.209}$$

for every $n\geq1$. Therefore, the final step of our proof will be to show that

$$\lim_{n\rightarrow\infty}\mathbf{E}\left[\phi\left(X_{n}\right)\right]=\mathbf{E}\left[\phi\left(X\right)\right]. \tag{5.210}$$

As we will accomplish these task, passing to the limit in (5.208), as n goes to infinity, and considering (5.208) and (5.210), the desired result will follow. □.

5.4 Spaces of Random Variables

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space and let $\mathcal{L}^0(\Omega; \mathbb{R})$ the linear space of all real $\mathcal{E}$-random variables on Ω.

Definition 659 *Given any $p > 0$, we say that $X \in \mathcal{L}^0(\Omega; \mathbb{R})$ has* finite moment of order p, *if the positive $\mathcal{E}$-random variable $|X|^p$ has finite expectation. We write $\mathcal{L}^p(\Omega; \mathbb{R})$ for the set of all real $\mathcal{E}$-random variables on Ω having finite moment of order p. Formally*

$$\mathcal{L}^p(\Omega; \mathbb{R}) \overset{def}{=} \left\{ X \in \mathcal{L}^0(\Omega; \mathbb{R}) : \int_\Omega |X|^p \, d\mathbf{P} < \infty \right\}.$$

Theorem 660 *Let $p \in (1, +\infty)$, with conjugate exponent q, let $X \in \mathcal{L}^p(\Omega; \mathbb{R})$ and let $Y \in \mathcal{L}^q(\Omega; \mathbb{R})$. Then, the random variable $XY \in \mathcal{L}^1(\Omega; \mathbb{R})$ and we have*

$$\mathbf{E}[|XY|] \leq \mathbf{E}[|X|^p]^{1/p} \mathbf{E}[|Y|^q]^{1/q}. \tag{5.211}$$

Proof. Write $A \equiv \mathbf{E}[|X|^p]^{1/p}$ and $B \equiv \mathbf{E}[|Y|^q]^{1/q}$ and assume that both A and B are different from zero. Otherwise, we would have $X = Y = 0$ a.s. on $\not\leqslant$ and (5.211) would be trivially true. Now, considering $x \equiv |X(\omega)| / A$ and $y \equiv |Y(\omega)| / B$, a straightforward application of Inequality (**??**) gives

$$\frac{|(XY)(\omega)|}{AB} \equiv \frac{|X(\omega) Y(\omega)|}{AB} \leq \frac{|X(\omega)|^p}{pA^p} + \frac{|Y(\omega)|^q}{qB^q}, \tag{5.212}$$

for every $\omega \in \Omega$. By virtue of the monotonicity property of the Lebesgue integral, Inequality (5.212) implies that the random variable XY has finite expectation. Moreover, computing the expectation of the first and third term of (5.212), we obtain

$$\frac{1}{AB}\mathbf{E}[|XY|] \leq \frac{1}{pA^p}\mathbf{E}[|X|^p] + \frac{1}{qB^q}\mathbf{E}[|Y|^q],$$

that is

$$\frac{\mathbf{E}[|XY|]}{\mathbf{E}[|X|^p]^{1/p} \mathbf{E}[|Y|^q]^{1/q}} \leq \frac{1}{p} + \frac{1}{q} = 1,$$

which clearly yields the Hölder inequality.**Proof.**

Corollary 661 (Cauchy-Schwarz) *Let $p \equiv 2$. Then, the conjugate esponent of p is $q = 2$ and for every $X, Y \in \mathcal{L}^2(\Omega; \mathbb{R})$ the random variable $XY \in \mathcal{L}^1(\Omega; \mathbb{R})$ and we have*

$$\mathbf{E}[|XY|] \leq \mathbf{E}\left[X^2\right]^{1/2} \mathbf{E}\left[Y^2\right]^{1/2}. \tag{5.213}$$

Theorem 662 (Minkowski inequality) *Let $p \in [1, +\infty)$ and let $X, Y \in \mathcal{L}^p(\Omega; \mathbb{R})$. Then, the random variable $X + Y \in \mathcal{L}^p(\Omega; \mathbb{R})$ and we have*

$$\mathbf{E}[|X + Y|^p]^{1/p} \leq \mathbf{E}[|X|^p]^{1/p} + \mathbf{E}[|Y|^p]^{1/p}. \tag{5.214}$$

Proof. Applying Lemma **??** we have

$$\frac{1}{2^p}\left|X\left(\omega\right)+Y\left(\omega\right)\right|^p \leq \frac{1}{2}\left(\left|X\left(\omega\right)\right|^p+\left|Y\left(\omega\right)\right|^p\right),$$

for every $\omega \in \nleqslant$. This yields the integrability of $|X+Y|^p$. Moreover, we can write

$$\mathbf{E}\left[|X+Y|^p\right] \leq \mathbf{E}\left[|X|\,|X+Y|^{p-1}\right]+\mathbf{E}\left[|Y|\,|X+Y|^{p-1}\right]. \tag{5.215}$$

Thus, applying the Hõlder inequality to both terms of the right hand side of (5.215), it follows

$$\mathbf{E}\left[|X+Y|^p\right] \leq \mathbf{E}\left[|X|^p\right]^{1/p}\mathbf{E}\left[|X+Y|^{q(p-1)}\right]^{1/q}+\mathbf{E}\left[|Y|^p\right]^{1/p}\mathbf{E}\left[|X+Y|^{q(p-1)}\right]^{1/q}. \tag{5.216}$$

Hence, under the obvious assumption $\mathbf{E}\left[|X+Y|^p\right] > 0$, dividing both sides of (5.216) by $\mathbf{E}\left[|X+Y|^{q(p-1)}\right]^{1/q}$, on account of $q(p-1)=p$ and $1-1/q=1/p$, we obtain the Minkowski inequality. □

Theorem 663 (Lyapunov inequality) *Assume $p \in [1,+\infty)$ and $X \in \mathcal{L}^p\left(\Omega;\mathbb{R}\right)$. Then, $X \in \mathcal{L}^\ell\left(\Omega;\mathbb{R}\right)$ for every $\ell \in [1,p]$. In addition, we have*

$$E\left[|X|^\ell\right]^{1/\ell} \leq E\left[|X|^p\right]^{1/p}. \tag{5.217}$$

Proof. Since the events $\{|X|<1\}$ and $\{|X|\geq 1\}$ constitute a partition of Ω, and $|X|^\ell \leq |X|^p$ on occurring the event $\{|X|\geq 1\}$, we can write

$$\begin{aligned}\int_\Omega |X|^\ell\, d\mathbf{P} &= \int_{\{|X|<1\}} |X|^\ell\, d\mathbf{P}+\int_{\{|X|\geq 1\}} |X|^\ell\, d\mathbf{P}\\ &\leq \int_{\{|X|<1\}} 1 d\mathbf{P}+\int_{\{|X|\geq 1\}} |X|^p\, d\mathbf{P}\\ &\leq \mathbf{P}\left(|X|<1\right)+\int_\Omega |X|^p\, d\mathbf{P}\\ &\leq 1+\int_\Omega |X|^p\, d\mathbf{P}.\end{aligned}$$

This proves the first claim. Regarding the second claim, which is not trivial for $p>1$, we know that $X \in \mathcal{L}^\ell\left(\Omega;\mathbb{R}\right)$, for every $\ell \in [1,p]$, the function $\varphi : \mathbb{R} \to \mathbb{R}$ given by

$$\varphi\left(x\right) \overset{\text{def}}{=} x^{p/\ell}, \quad \forall x \in \mathbb{R},$$

is convex and $\varphi\left(|X|^\ell\right) = |X|^p \in \mathcal{L}^1\left(\Omega;\mathbb{R}\right)$. Hence, we can apply the Jensen inequality to the random variable $|X|^\ell \in \mathcal{L}^1\left(\Omega;\mathbb{R}\right)$ obtaining

$$\varphi\left(\mathbf{E}\left[|X|^\ell\right]\right) \leq \mathbf{E}\left[\varphi\left(|X|^\ell\right)\right],$$

that is

$$\mathbf{E}\left[|X|^\ell\right]^{p/\ell} \leq \mathbf{E}\left[|X|^p\right].$$

From the latter, Equation (5.217) immediately follows. Another proof of the second claim can be given by using the Hölder inequality. In fact, considering the random variables $|X|^{\ell} \in \mathcal{L}^{p/\ell}(\Omega;\mathbb{R})$ and $1_{\Omega} \in \mathcal{L}^{p/(p-\ell)}(\Omega;\mathbb{R})$ and applying the Hölder inequality with conjugate exponents p/ℓ and $p/(p-\ell)$, we can write

$$\mathbf{E}\left[\left||X|^{\ell}1_{\Omega}\right|\right] \leq \mathbf{E}\left[\left||X|^{\ell}\right|^{p/\ell}\right]^{\ell/p} \mathbf{E}[1_{\Omega}^{p/(p-\ell)}]^{(p-\ell)/p} = \mathbf{E}\left[|X|^{p}\right]^{\ell/p} \mathbf{E}[1_{\Omega}]^{(p-\ell)/p} = \left(\mathbf{E}\left[|X|^{p}\right]^{1/p}\right)^{\ell}.$$

It then follows

$$\mathbf{E}\left[|X|^{\ell}\right] = \mathbf{E}\left[\left||X|^{\ell}1_{\Omega}\right|\right] \leq \left(\mathbf{E}\left[|X|^{p}\right]^{1/p}\right)^{\ell},$$

which implies Equation (5.217). □

5.4.1 Covariance and Correlation

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space, let $\mathcal{L}^0(\Omega;\mathbb{R})$ the linear space of all real $\mathcal{E}$-random variables on Ω, and let $\mathcal{L}^2(\Omega;\mathbb{R})$ be the subset of $\mathcal{L}^0(\Omega;\mathbb{R})$ containing all random variables having finite moment of order 2.

Definition 664 *For all $X, Y \in \mathcal{L}^2(\Omega;\mathbb{R})$, we call the* covariance *of X and Y the real number*

$$Cov(X,Y) \overset{def}{=} \mathbf{E}\left[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])\right]. \tag{5.218}$$

Note that for all $X, Y \in \mathcal{L}^2(\Omega;\mathbb{R})$ also $X - \mathbf{E}[X], Y - \mathbf{E}[Y] \in \mathcal{L}^2(\Omega;\mathbb{R})$ (see Corollary 588). Hence, the product $(X - \mathbf{E}[X])(Y - \mathbf{E}[Y]) \in \mathcal{L}^1(\Omega;\mathbb{R})$. It follows $Cov(X,Y) \in \mathbb{R}$. Note also that, since the events $\{(X - \mathbf{E}[X])(Y - \mathbf{E}[Y]) \geq 0\} \equiv C$ and $\{(X - \mathbf{E}[X])(Y - \mathbf{E}[Y]) < 0\} \equiv D$ costitute a partition of Ω, we can write

$$Cov(X,Y) = \int_C (X - \mathbf{E}[X])(Y - \mathbf{E}[Y])\, d\mathbf{P} + \int_D (X - \mathbf{E}[X])(Y - \mathbf{E}[Y])\, d\mathbf{P},$$

where

$$\int_C (X - \mathbf{E}[X])(Y - \mathbf{E}[Y])\, d\mathbf{P} \geq 0 \quad \text{and} \quad \int_D (X - \mathbf{E}[X])(Y - \mathbf{E}[Y])\, d\mathbf{P} \leq 0.$$

Therefore, when $Cov(X,Y) > 0$ [resp. $Cov(X,Y) < 0$] the concordant [resp. discordant] deviations of X and Y from the expectations $\mathbf{E}[X]$ and $\mathbf{E}[Y]$, averaged on the event where they are concordant [resp. discordant], prevail over the discordant [resp. concordant] deviations, averaged on the event where they are discordant [resp. concordant].

Remark 665 *We have*

$$Cov(X,X) = \mathbf{D}^2[X], \tag{5.219}$$

for every $X \in \mathcal{L}^2(\Omega;\mathbb{R})$.

Proposition 666 *We have*

$$Cov(X,Y) = \mathbf{E}[XY] - \mathbf{E}[X]\,\mathbf{E}[Y], \tag{5.220}$$

for all $X, Y \in \mathcal{L}^2(\Omega;\mathbb{R})$.

Proof. Applying the properties of the expectation operator, we obtain

$$\begin{aligned}\mathbf{E}\left[\left(X-\mathbf{E}\left[X\right]\right)\left(Y-\mathbf{E}\left[Y\right]\right)\right] &= \mathbf{E}\left[XY-X\mathbf{E}\left[Y\right]-\mathbf{E}\left[X\right]Y+\mathbf{E}\left[X\right]\mathbf{E}\left[Y\right]\right]\\ &= \mathbf{E}\left[XY\right]-\mathbf{E}\left[X\mathbf{E}\left[Y\right]\right]-\mathbf{E}\left[\mathbf{E}\left[X\right]Y\right]+\mathbf{E}\left[\mathbf{E}\left[X\right]\mathbf{E}\left[Y\right]\right]\\ &= \mathbf{E}\left[XY\right]-\mathbf{E}\left[X\right]\mathbf{E}\left[Y\right]-\mathbf{E}\left[X\right]\mathbf{E}\left[Y\right]+\mathbf{E}\left[X\right]\mathbf{E}\left[Y\right]\\ &= \mathbf{E}\left[XY\right]-\mathbf{E}\left[X\right]\mathbf{E}\left[Y\right],\end{aligned}$$

which is the desired (5.220). □

Proposition 667 *We have*

1. $Cov\left(X,Y\right)=Cov\left(Y,X\right)$, *for all* $X,Y\in\mathcal{L}^2\left(\Omega;\mathbb{R}\right)$;
2. $Cov\left(\alpha X+\beta Y,Z\right)=\alpha Cov\left(X,Z\right)+\beta Cov\left(Y,Z\right)$ *for all* $X,Y,Z\in\mathcal{L}^2\left(\Omega;\mathbb{R}\right)$.

Otherwise saying, the functional $Cov\left(\cdot,\cdot\right):\mathcal{L}^2\left(\Omega;\mathbb{R}\right)\times\mathcal{L}^2\left(\Omega;\mathbb{R}\right)\to\mathbb{R}$ *given by Equation (5.218) is bilinear and symmetric.*

Proof. . □

Proposition 668 *We have*

$$\left|Cov\left(X,Y\right)\right|\leq\mathbf{D}\left[X\right]\mathbf{D}\left[Y\right], \tag{5.221}$$

for all $X,Y\in\mathcal{L}^2\left(\Omega;\mathbb{R}\right)$.

Proof. Applying the Cauchy-Schwarz inequality (5.213) to the random variables $X-\mathbf{E}\left[X\right]$ and $Y-\mathbf{E}\left[Y\right]$ we obtain

$$\mathbf{E}\left[\left|\left(X-\mathbf{E}\left[X\right]\right)\left(Y-\mathbf{E}\left[Y\right]\right)\right|\right]\leq\mathbf{E}\left[\left(X-\mathbf{E}\left[X\right]\right)^2\right]^{1/2}\mathbf{E}\left[\left(Y-\mathbf{E}\left[Y\right]\right)^2\right]^{1/2}=\mathbf{D}\left[X\right]\mathbf{D}\left[Y\right]. \tag{5.222}$$

On the other hand,

$$\left|Cov\left(X,Y\right)\right|=\left|\mathbf{E}\left[\left(X-\mathbf{E}\left[X\right]\right)\left(Y-\mathbf{E}\left[Y\right]\right)\right]\right|\leq\mathbf{E}\left[\left|\left(X-\mathbf{E}\left[X\right]\right)\left(Y-\mathbf{E}\left[Y\right]\right)\right|\right]. \tag{5.223}$$

Combining (5.222) and (5.223), the desired (5.221) immediately follows.

Definition 669 *For all* $X,Y\in\mathcal{L}^2\left(\Omega;\mathbb{R}\right)$ *we call the* correlation *of* X *and* Y *the real number*

$$Corr\left(X,Y\right)\overset{def}{=}\begin{cases}\frac{Cov(X,Y)}{\mathbf{D}[X]\mathbf{D}[Y]}, & \text{if } \mathbf{D}\left[X\right]\mathbf{D}\left[Y\right]>0,\\ 0, & \text{if } \mathbf{D}\left[X\right]\mathbf{D}\left[Y\right]=0.\end{cases}$$

Remark 670 *We have*

$$Corr\left(X,Y\right)=0\Leftrightarrow\mathbf{E}\left[XY\right]=\mathbf{E}\left[X\right]\mathbf{E}\left[Y\right].$$

Proposition 671 *We have*

$$\left|Corr\left(X,Y\right)\right|\leq 1. \tag{5.224}$$

for all $X,Y\in\mathcal{L}^2\left(\Omega;\mathbb{R}\right)$.

Proof. Equation (5.221) clearly implies Equation (**??**) in both cases $\mathbf{D}[X]\mathbf{D}[Y] = 0$ and $\mathbf{D}[X]\mathbf{D}[Y] > 0$. □

Definition 672 *Given any pair $X, Y \in \mathcal{L}^2(\Omega;\mathbb{R})$, we say that X, Y are* perfectly correlated *[resp.* perfectly anticorrelated*, resp.* uncorrelated*] if $Corr(X,Y) = 1$ [resp. $Corr(X,Y) = -1$, resp. $Corr(X,Y) = 0$].*

Definition 673 *We say that X and Y are* orthogonal *if*

$$\mathbf{E}[XY] = 0.$$

Remark 674 *The random variables X and Y are uncorrelated if and only if the random variables $X - \mathbf{E}[X]$ and $Y - \mathbf{E}[Y]$ are ortogonal.*

5.4.2 Linear Spaces $\mathcal{L}^p(\Omega;\mathbb{R})$

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space, let $\mathcal{L}^0(\Omega;\mathbb{R})$ the linear space of all real $\mathcal{E}$-random variables on Ω, and let $\mathcal{L}^p(\Omega;\mathbb{R})$ be the subset of $\mathcal{L}^0(\Omega;\mathbb{R})$ containing all random variables having finite moment of order p.

Proposition 675 *Let $1 \leq p < \infty$. Then, the set $\mathcal{L}^p(\Omega;\mathbb{R})$ is a real linear space and the map $|\cdot|_p : \mathcal{L}^p(\Omega;\mathbb{R}) \to \mathbb{R}_+$, given by*

$$|X|_p \overset{def}{=} \mathbf{E}[|X|^p]^{1/p}, \quad \forall X \in \mathcal{L}^p(\Omega;\mathbb{R}), \tag{5.225}$$

is a seminorm.

Proof. The set $\mathcal{L}^p(\Omega;\mathbb{R})$ is non-empty since it clearly contains the Dirac random variables. In addition, by virtue of the Minkowski inequality and the homogeneity property of the expectation, we can write

$$\mathbf{E}[|\alpha X + \beta Y|^p]^{1/p} \leq \mathbf{E}[|\alpha X|^p]^{1/p} + \mathbf{E}[|\beta Y|^p]^{1/p} = |\alpha|\,\mathbf{E}[|X|^p]^{1/p} + |\beta|\,\mathbf{E}[|Y|^p]^{1/p}. \tag{5.226}$$

for all $X, Y \in \mathcal{L}^p(\Omega;\mathbb{R})$ and all $\alpha, \beta \in \mathbb{R}$. It immediately follows that $\mathcal{L}^p(\Omega;\mathbb{R})$ is a linear submanifold of $\mathcal{RV}(\Omega;\mathbb{R})$. Hence, a real linear space. Moreover, combining (5.225) and (5.226), we obtain

$$|\alpha X + \beta Y|_p \leq |\alpha|\,|X|_p + |\beta|\,|Y|_p,$$

for all $\alpha, \beta \in \mathbb{R}$ and all $X, Y \in \mathcal{L}^p(\Omega;\mathbb{R})$, which shows that the map $|\cdot|_p$ is a seminorm. □

Remark 676 *The map $|\cdot|_p : \mathcal{L}^p(\Omega;\mathbb{C}) \to \mathbb{R}_+$ is not a norm because*

$$|X|_p = 0 \nRightarrow X = 0.$$

However,

$$|X|_p = 0 \Rightarrow X = 0 \quad \mathbf{P}\text{-}a.s.\ on\ \Omega.$$

Definition 677 *We call the seminorm given by Equation (5.225) the $\mathcal{L}^p$-seminorm.*

Theorem 678 *Let $(X_n)_{n\geq 1}$ be a Cauchy sequence in $\mathcal{L}^p(\Omega;\mathbb{R})$. Then, there exists $X \in \mathcal{L}^p(\Omega;\mathbb{R})$ such that*

$$\lim_{n\to\infty} |X_n - X|_p = 0. \tag{5.227}$$

In addition, there exists a subsequence $(X_{n_k})_{k\geq 1}$ of $(X_n)_{n\geq 1}$ which converges to X almost surely on Ω.

Proof. Since $(X_n)_{n\geq 1}$ is a Cauchy sequence in $\mathcal{L}^p(\Omega;\mathbb{R})$, we can extract a subsequence $(X_{n_k})_{k\geq 1}$ such that

$$\left|X_{n_k} - X_{n_{k-1}}\right|_p < 2^{-k},$$

for every $k \geq 1$. Hence, considering the nondecreasing sequence $(Y_n)_{n\geq 1}$ of positive random variables on Ω given by

$$Y_n \overset{\text{def}}{=} \sum_{k=1}^{n} \left|X_{n_k} - X_{n_{k-1}}\right|, \quad \forall n \geq 1,$$

an iterated application of the Minkowski inequality (5.214) yields

$$|Y_n|_p = \mathbf{E}\left[\left|\sum_{k=1}^{n} \left|X_{n_k} - X_{n_{k-1}}\right|\right|^p\right]^{1/p} \leq \sum_{k=1}^{n} \mathbf{E}\left[\left|X_{n_k} - X_{n_{k-1}}\right|^p\right]^{1/p} = \sum_{k=1}^{n} \left|X_{n_k} - X_{n_{k-1}}\right|_p \leq \sum_{k=1}^{n} 2^{-k} = 1.$$

for every $n \geq 1$. Furthermore, the random variable $Y : \Omega \to \overline{\mathbb{R}}_+$ given by

$$Y(\omega) \overset{\text{def}}{=} \lim_{n\to\infty} Y_n(\omega), \quad \forall \omega \in \Omega,$$

is well defined. Now, by virtue of the Fatou lemma, we can write

$$\int_\Omega Y^p \, d\mathbf{P} \leq \liminf_{n\to\infty} \int_\Omega Y_n^p \, d\mathbf{P} = \liminf_{n\to\infty} |Y_n|_p^p \leq 1.$$

This implies that we have $Y^p < \infty$ almost surely on Ω. It follows that the series

$$\sum_{k=1}^{\infty} \left(X_{n_k}(\omega) - X_{n_{k-1}}(\omega)\right)$$

converges absolutely almost surely on Ω and it is well defined the real random variable X on Ω given by

$$X(\omega) \overset{\text{def}}{=} \begin{cases} \lim_{n\to\infty} X_{n_1}(\omega) + \sum_{k=1}^{n} \left(X_{n_k}(\omega) - X_{n_{k-1}}(\omega)\right) & \text{if } Y(\omega) < \infty \\ 0 & \text{otherwise} \end{cases}.$$

In addition, since

$$X_{n_1}(\omega) + \sum_{k=1}^{n} \left(X_{n_k}(\omega) - X_{n_{k-1}}(\omega)\right) = X_{n_n}(\omega),$$

for every $n \geq 1$, the above arguments prove that the sequence $(X_{n_k})_{k\geq 1}$ eventually converges almost surely on Ω to the random variable X. We are left with showing that X belongs to $\mathcal{L}^p(\Omega;\mathbb{R})$ and Condition (5.227) holds true. In fact, for any fixed $\varepsilon > 0$ there exists $n_\varepsilon \geq 1$ such that

$$|X_n - X_m|_p < \varepsilon,$$

for all $n, m \geq n_\varepsilon$. Therefore, for any fixed $m \geq n_\varepsilon$, applying again the Fatou lemma, we have

$$\int_\Omega |X - X_m|^p \, d\mathbf{P} \leq \liminf_{n\to\infty} \int_\Omega |X_n - X_m|^p \, \mathbf{P}(d\omega) = \liminf_{n\to\infty} |X_n - X_m|_p^p < \varepsilon^p.$$

Hence,

$$\int_\Omega |X|^p \, d\mathbf{P} \leq \int_\Omega |X - X_m|^p \, d\mathbf{P} + \int_\Omega |X_m|^p \, d\mathbf{P} \leq \varepsilon^p + |X_m|_p^p < \infty,$$

which implies $X \in \mathcal{L}^p(\Omega; \mathbb{R})$. Finally, we can write

$$|X - X_n|_p \leq |X - X_m|_p + |X_n - X_m|_p < 2\varepsilon,$$

and this proves the convergence of $(X_n)_{n\geq 1}$ to X in $\mathcal{L}^p(\Omega; \mathbb{R})$. □

5.4.3 Banach spaces $L^p(\Omega; \mathbb{R})$

For any $1 \leq p < \infty$ let $\mathcal{L}^p(\Omega; \mathbb{R})$ be the real linear space of all real $\mathcal{E}$-random variables on Ω having finite moment of order p and let $|\cdot|_p : \mathcal{L}^p(\Omega; \mathbb{R}) \to \mathbb{R}_+$ the $\mathcal{L}^p$-seminorm (see Definition **??**).

Proposition 679 *The seminorm $|\cdot|_p : \mathcal{L}^p(\Omega; \mathbb{R}) \to \mathbb{R}_+$ defines an equivalence relation on $\mathcal{L}^p(\Omega; \mathbb{R})$ given by*

$$X \sim Y \bmod |\cdot|_p \Leftrightarrow |X - Y|_p = 0, \quad \forall X, Y \in \mathcal{L}^p(\Omega; \mathbb{R}).$$

We have

$$X \sim Y \bmod |\cdot|_p \Leftrightarrow X \overset{\mathbf{P}\text{-}a.s.}{=} Y,$$

and the equivalence relation preserves the linear structure of $\mathcal{L}^p(\Omega; \mathbb{R})$.

Proof. . □

Notation 680 *We write $L^p(\Omega; \mathbb{R})$ for the quotient space $\mathcal{L}^p(\Omega; \mathbb{R}) / \sim \bmod |\cdot|_p$.*

In light of what shown above it follows

Theorem 681 *The space $L^p(\Omega; \mathbb{R})$ is a Banach space with respect to the norm $\|\cdot\|_p : L^p(\Omega; \mathbb{R}) \to \mathbb{R}_+$ given by*

$$\left\|\tilde{X}\right\|_p \overset{def}{=} |X|_p, \quad \forall \tilde{X} \in L^p(\Omega; \mathbb{R}),$$

where $X \in \mathcal{L}^p(\Omega; \mathbb{R})$ is any representative of the equivalence class $\tilde{X}$.

Proof. . □

Notation 682 *When no confusion can arise, it is customary to denote the equivalence classes $\tilde{X}, \tilde{Y}, \tilde{Z}, \ldots$ in $L^p(\Omega; \mathbb{R})$ of the random variables $X, Y, Z, \ldots$ in $\mathcal{L}^p(\Omega; \mathbb{R})$ dropping the "hat" $\sim$ and using the same capital letters $X, Y, Z, \ldots$ which denote their representatives.*

5.4.4 Hilbert space $L^2(\Omega;\mathbb{R})$

Let $L^2(\Omega;\mathbb{R})$ the Banach space of all real $\mathcal{E}$-random variables having finite moment of order 2.

Theorem 683 *The map* $\langle\cdot,\cdot\rangle_2 : L^2(\Omega;\mathbb{R}) \times L^2(\Omega;\mathbb{R}) \to \mathbb{C}$ *given by*

$$\left\langle \tilde{X}, \tilde{Y} \right\rangle_2 \overset{def}{=} \mathbf{E}[XY], \quad \forall \tilde{X}, \tilde{Y} \in L^2(\Omega;\mathbb{R}),$$

where $X \in \tilde{X}$ *and* $Y \in \tilde{Y}$ *is a scalar product on* $L^2(\Omega;\mathbb{R})$ *and the norm associated to* $\langle\cdot,\cdot\rangle_2$ *is the* L^2*-norm.*

Proof. . □

Corollary 684 *The space* $L^2(\Omega;\mathbb{R})$ *is a Hilbert space with respect to the scalar product* $\langle\cdot,\cdot\rangle_2$.

Let $\tilde{X}, \tilde{Y} \in L^2(\Omega;\mathbb{R})$.

Definition 685 *According to the Hilbert spaces terminology we say that* $\tilde{X}$ *and* $\tilde{Y}$ *are* orthogonal *if*

$$\left\langle \tilde{X}, \tilde{Y} \right\rangle_2 = 0.$$

Remark 686 *The equivalence classes* $\tilde{X}, \tilde{Y}$ *are orthogonal if and only if*

$$\mathbf{E}[XY] = 0$$

for every $X \in \tilde{X}$ *and* $Y \in \tilde{Y}$.

Chapter 6

Real Random Vectors

6.1 Basic Definitions and Notation

Fixed any $N \in \mathbb{N}$, let $\left(\mathbb{R}^N, \mathcal{B}(\mathbb{R}^N), \mu_L^N\right) \equiv \mathbb{R}^N$ be the real N-dimensional Euclidean space equipped with the Borel σ-algebra $\mathcal{B}(\mathbb{R}^N)$ and the Borel-Lebesgue measure $\mu_L^N : \mathcal{B}(\mathbb{R}^N) \to \mathbb{R}_+$. Recall that we adopt the standard convention of identifying a point $x \in \mathbb{R}^N$ with the column vector $(x_1, \ldots, x_N)^\intercal$ of its entries (see 1). Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space and let $X : \Omega \to \mathbb{R}^N$ be a map from Ω to $\mathbb{R}^N$, denoted briefly by X when no confusion can arise.

Definition 687 *For any $K = 1, \ldots, N$, we call the Kth* entry *of X the real function $X_K : \Omega \to \mathbb{R}$, briefly X_K, given by*

$$X_K \overset{def}{=} \pi_K \circ X, \tag{6.1}$$

where $\pi_K : \mathbb{R}^N \to \mathbb{R}$ is the Kth canonical projection of $\mathbb{R}^N$ over $\mathbb{R}$ (see Equation (1.4)).

Let $X_1 : \Omega \to \mathbb{R}, \ldots, X_N : \Omega \to \mathbb{R}$, briefly $X_1, \ldots, X_N$, be the entries of X.

Remark 688 *We have*

$$X_K(\omega) = \pi_K(X(\omega)) = \pi_K(x) = x_K, \tag{6.2}$$

for any $n = 1, \ldots, N$, and

$$X(\omega) = (X_1(\omega), \ldots, X_N(\omega))^\intercal, \tag{6.3}$$

where $X(\omega) \equiv x \equiv (x_1, \ldots, x_N)^\intercal$, for every $\omega \in \Omega$.

Notation 689 *Considering Equation (6.3), to express that $X_1, \ldots, X_N$ are the entries of X it is customary to write $X \equiv (X_1, \ldots, X_N)^\intercal$.*

Generalizing Definition 402, we can state

Definition 690 *We say that the map $X : \Omega \to \mathbb{R}^N$ is an N-*dimensional real $\mathcal{E}$-random vector *or an N-v*ariate real $\mathcal{E}$-random variable *on Ω or an $\mathcal{E}$-*random variable *on Ω with states in $\mathbb{R}^N$, if the X-inverse image of any Borel set in $\mathcal{B}\left(\mathbb{R}^N\right)$ is an event in $\mathcal{E}$. In symbols,*

$$\{X \in B\} \in \mathcal{E}, \quad \forall B \in \mathcal{B}\left(\mathbb{R}^N\right). \tag{6.4}$$

Corresponding to Proposition 1421 we have

Proposition 691 *Assume $\mathcal{B}$ is a basis for $\mathcal{B}\left(\mathbb{R}^N\right)$, that is $\sigma\left(\mathcal{B}\right) = \mathcal{B}\left(\mathbb{R}^N\right)$. Then, the map $X : \Omega \to \mathbb{R}^N$ is an $\mathcal{E}$-random vector if and only the X-inverse image of any Borel set in $\mathcal{B}$ is an event in $\mathcal{E}$. In symbols,*

$$\{X \in B\} \in \mathcal{E}, \quad \forall B \in \mathcal{B}.$$

Remark 692 *The map $X : \Omega \to \mathbb{R}^N$ is an $\mathcal{E}$-random vector if and only if*

$$\{X \in I\} \in \mathcal{E}, \quad \forall I \in \mathcal{I},$$

where $\mathcal{I}$ is any of the families $\mathcal{I}_{o.\mathbb{R}}\left(\mathbb{R}^N\right)$, $\mathcal{I}_{c.\mathbb{R}}\left(\mathbb{R}^N\right)$, $\mathcal{I}_{o.c.\mathbb{R}}\left(\mathbb{R}^N\right)$, $\mathcal{I}_{c.o.\mathbb{R}}\left(\mathbb{R}^N\right)$ introduced in Proposition 49 or any of the families $\mathcal{I}_{o.\mathbb{Q}}\left(\mathbb{R}^N\right)$, $\mathcal{I}_{c.\mathbb{Q}}\left(\mathbb{R}^N\right)$, $\mathcal{I}_{o.c.\mathbb{Q}}\left(\mathbb{R}^N\right)$, $\mathcal{I}_{c.o.\mathbb{Q}}\left(\mathbb{R}^N\right)$ introduced in Proposition 50.

Analogously,

Remark 693 *The map $X : \Omega \to \mathbb{R}^N$ is an $\mathcal{E}$-random vector if and only if*

$$\{X \in H\} \in \mathcal{E}, \quad \forall H \in \mathcal{H},$$

where $\mathcal{H}$ is any of the families $\mathcal{H}_{r.o.\mathbb{R}}\left(\mathbb{R}^N\right)$, $\mathcal{H}_{r.c.\mathbb{R}}\left(\mathbb{R}^N\right)$, $\mathcal{H}_{l.o.\mathbb{R}}\left(\mathbb{R}^N\right)$, $\mathcal{H}_{l.c.\mathbb{R}}\left(\mathbb{R}^N\right)$ introduced in Proposition 51 or any of the families $\mathcal{H}_{r.o.\mathbb{Q}}\left(\mathbb{R}^N\right)$, $\mathcal{H}_{r.c.\mathbb{Q}}\left(\mathbb{R}^N\right)$, $\mathcal{H}_{l.o.\mathbb{Q}}\left(\mathbb{R}^N\right)$, $\mathcal{H}_{l.c.\mathbb{Q}}\left(\mathbb{R}^N\right)$ introduced in Proposition 52.

Recalling the notation introduced in Proposition 49 [resp. 50], we can write $I \equiv \mathsf{X}_{K=1}^{N} I_K$, where $I_K \equiv (a_K, b_K)$ or $I_K \equiv [a_K, b_K]$ or $I_K \equiv (a_K, b_K]$ or $I_K \equiv [a_K, b_K)$, for some $(a_1, \ldots, a_N)$, $(b_1, \ldots, b_N) \in \mathbb{R}^N$ [resp. $(a_1, \ldots, a_N)$, $(b_1, \ldots, b_N) \in \mathbb{Q}^N$], according to whether I is in $\mathcal{I}_{o.\mathbb{R}}\left(\mathbb{R}^N\right)$ or $\mathcal{I}_{c.\mathbb{R}}\left(\mathbb{R}^N\right)$ or $\mathcal{I}_{o.c.\mathbb{Q}}\left(\mathbb{R}^N\right)$ or $\mathcal{I}_{c.o.\mathbb{Q}}\left(\mathbb{R}^N\right)$ [resp. $\mathcal{I}_{o.\mathbb{Q}}\left(\mathbb{R}^N\right)$ or $\mathcal{I}_{c.\mathbb{Q}}\left(\mathbb{R}^N\right)$ or $\mathcal{I}_{o.c.\mathbb{Q}}\left(\mathbb{R}^N\right)$ or $\mathcal{I}_{c.o.\mathbb{Q}}\left(\mathbb{R}^N\right)$]. Similarly, recalling the notation introduced in Proposition 51 [resp. 52] we can write $H \equiv \mathsf{X}_{K=1}^{N} H_K$, where $H_K \equiv (-\infty, a_K)$ or $H_K \equiv (-\infty, a_K]$ or $H_K \equiv (a_K, +\infty)$ or $H_K \equiv [a_K, +\infty)$, for some $(a_1, \ldots, a_N) \in \mathbb{R}^N$ [resp. $(a_1, \ldots, a_N) \in \mathbb{Q}^N$], according to whether H is in $\mathcal{H}_{r.o.\mathbb{R}}\left(\mathbb{R}^N\right)$ or $\mathcal{H}_{r.c.\mathbb{R}}\left(\mathbb{R}^N\right)$ or $\mathcal{H}_{l.o.\mathbb{R}}\left(\mathbb{R}^N\right)$ or $\mathcal{H}_{l.c.\mathbb{R}}\left(\mathbb{R}^N\right)$ [resp. $\mathcal{H}_{r.o.\mathbb{Q}}\left(\mathbb{R}^N\right)$ or $\mathcal{H}_{r.c.\mathbb{Q}}\left(\mathbb{R}^N\right)$ or $\mathcal{H}_{l.o.\mathbb{Q}}\left(\mathbb{R}^N\right)$ or $\mathcal{H}_{l.c.\mathbb{Q}}\left(\mathbb{R}^N\right)$]. Recalling also that $\{X \in I\}$ [resp. $\{X \in H\}$] is a shorthand for $\{\omega \in \Omega : X(\omega) \in I\}$ [resp. $\{\omega \in \Omega : X(\omega) \in H\}$], in terms of the entries $X_1, \ldots, X_N$ of X, we can write

$$\begin{aligned} \{X \in I\} &\equiv \left\{\omega \in \Omega : \left(X_1(\omega), \ldots, X_N(\omega)\right)^{\intercal} \in \mathsf{X}_{K=1}^{N} I_K\right\} \\ &\equiv \{\omega \in \Omega : X_1(\omega) \in I_1, \ldots, X_N(\omega) \in H_N\} \end{aligned}$$

$$\begin{aligned} [\text{resp. } \{X \in H\} &\equiv \left\{\omega \in \Omega : \left(X_1(\omega), \ldots, X_N(\omega)\right)^{\intercal} \in \mathsf{X}_{K=1}^{N} H_K\right\} \\ &\equiv \{\omega \in \Omega : X_1(\omega) \in H_1, \ldots, X_N(\omega) \in H_N\}\,]. \end{aligned}$$

Remark 694 *We have*

$$\{\omega \in \Omega : X_1(\omega) \in I_1, \ldots, X_N(\omega) \in I_N\} = \cap_{K=1}^{N} \{\omega \in \Omega : X_K(\omega) \in I_K\}. \tag{6.5}$$

for any $\mathsf{X}_{K=1}^{N} I_K \in \mathcal{I}$, where $\mathcal{I}$ is any of the families $\mathcal{I}_{o.\mathbb{R}}\left(\mathbb{R}^N\right)$, $\mathcal{I}_{c.\mathbb{R}}\left(\mathbb{R}^N\right)$, $\mathcal{I}_{o.c.\mathbb{R}}\left(\mathbb{R}^N\right)$, $\mathcal{I}_{c.o.\mathbb{R}}\left(\mathbb{R}^N\right)$ or $\mathcal{I}_{o.\mathbb{Q}}\left(\mathbb{R}^N\right)$, $\mathcal{I}_{c.\mathbb{Q}}\left(\mathbb{R}^N\right)$, $\mathcal{I}_{o.c.\mathbb{Q}}\left(\mathbb{R}^N\right)$, $\mathcal{I}_{c.o.\mathbb{Q}}\left(\mathbb{R}^N\right)$. Similarly,

$$\{\omega \in \Omega : X_1(\omega) \in H_1, \ldots, X_N(\omega) \in H_N\} = \cap_{K=1}^{N} \{\omega \in \Omega : X_K(\omega) \in H_K\}. \tag{6.6}$$

for any $\mathsf{X}_{K=1}^{N} H_K \in \mathcal{H}$, where $\mathcal{H}$ is any of the families $\mathcal{H}_{r.o.\mathbb{R}}\left(\mathbb{R}^N\right)$, $\mathcal{H}_{r.c.\mathbb{R}}\left(\mathbb{R}^N\right)$, $\mathcal{H}_{l.o.\mathbb{R}}\left(\mathbb{R}^N\right)$, $\mathcal{H}_{l.c.\mathbb{R}}\left(\mathbb{R}^N\right)$ or $\mathcal{H}_{r.o.\mathbb{Q}}\left(\mathbb{R}^N\right)$, $\mathcal{H}_{r.c.\mathbb{Q}}\left(\mathbb{R}^N\right)$, $\mathcal{H}_{l.o.\mathbb{Q}}\left(\mathbb{R}^N\right)$, $\mathcal{H}_{l.c.\mathbb{Q}}\left(\mathbb{R}^N\right)$.

Notation 695 *For the event in Equation (6.5) we will also use the standard shorthands*

$$\{a < X < b\} \equiv \{a_1 < X_1 < b_1, \ldots, a_N < X_N < b_N\} \tag{6.7}$$

or

$$\{a \leq X \leq b\} \equiv \{a_1 \leq X_1 \leq b_1, \ldots, a_N \leq X_N \leq b_N\} \tag{6.8}$$

or

$$\{a < X \leq b\} \equiv \{a_1 < X_1 \leq b_1, \ldots, a_N < X_N \leq b_N\} \tag{6.9}$$

or

$$\{a \leq X < b\} \equiv \{a_1 \leq X_1 < b_1, \ldots, a_N \leq X_N < b_N\} \tag{6.10}$$

according to whether $I_K \equiv (a_K, b_K)$ *or* $I_K \equiv [a_K, b_K]$ *or* $I_K \equiv (a_K, b_K]$ *or* $I_K \equiv [a_K, b_K)$*, for every* $K = 1, \ldots, N$*, and some* $a \equiv (a_1, \ldots, a_N)^\intercal$*,* $b \equiv (b_1, \ldots, b_N)^\intercal \in \mathbb{R}^N$*. Similarly, for the event in Equation (6.6), we will also use the standard shorthands*

$$\{X < a\} \equiv \{X_1 < a_1, \ldots, X_N < a_N\} \tag{6.11}$$

or

$$\{X \leq a\} \equiv \{X_1 \leq a_1, \ldots, X_N \leq a_N\} \tag{6.12}$$

or

$$\{X > a\} \equiv \{X_1 > a_1, \ldots, X_N > a_N\} \tag{6.13}$$

or

$$\{X \geq a\} \equiv \{X_1 \geq a_1, \ldots, X_N \geq a_N\} \tag{6.14}$$

according to whether $H_K \equiv (-\infty, a_K)$ *or* $H_K \equiv (-\infty, a_K]$ *or* $H_K \equiv (a_K, +\infty)$ *or* $H_K \equiv [a_K, +\infty)$*, for every* $K = 1, \ldots, N$*, and some* $a \equiv (a_1, \ldots, a_N)^\intercal \in \mathbb{R}^N$*.*

In terms of the notation introduced above, we have

Remark 696 *The map* $X : \Omega \to \mathbb{R}^N$ *is an* $\mathcal{E}$*-random vector if and only if*

$$\{a < X < b\} \equiv \{a_1 < X_1 < b_1, \ldots, a_N < X_N < b_N\} \in \mathcal{E} \tag{6.15}$$

or, equivalently,

$$\{a \leq X \leq b\} \equiv \{a_1 \leq X_1 \leq b_1, \ldots, a_N \leq X_N \leq b_N\} \in \mathcal{E} \tag{6.16}$$

or, equivalently,

$$\{a < X \leq b\} \equiv \{a_1 < X_1 \leq b_1, \ldots, a_N < X_N \leq b_N\} \in \mathcal{E} \tag{6.17}$$

or, equivalently,

$$\{a \leq X < b\} \equiv \{a_1 \leq X_1 < b_1, \ldots, a_N \leq X_N < b_N\} \in \mathcal{E} \tag{6.18}$$

for all $a \equiv (a_1, \ldots, a_N)^\intercal$*,* $b \equiv (b_1, \ldots, b_N)^\intercal \in \mathbb{R}^N$ *or* $a \equiv (a_1, \ldots, a_N)^\intercal$*,* $b \equiv (b_1, \ldots, b_N)^\intercal \in \mathbb{Q}^N$*. Similarly,* $X : \Omega \to \mathbb{R}^N$ *is an* $\mathcal{E}$*-random vector if and only if*

$$\{X < a\} \equiv \{X_1 < a_1, \ldots, X_N < a_N\} \in \mathcal{E} \tag{6.19}$$

or, equivalently,

$$\{X \leq a\} \equiv \{X_1 \leq a_1, \ldots, X_N \leq a_N\} \in \mathcal{E} \tag{6.20}$$

or, equivalently,

$$\{X > a\} \equiv \{X_1 > a_1, \ldots, X_N > a_N\} \in \mathcal{E} \tag{6.21}$$

or, equivalently,

$$\{X \geq a\} \equiv \{X_1 \geq a_1, \ldots, X_N \geq a_N\} \in \mathcal{E}, \tag{6.22}$$

for every $a \equiv (a_1, \ldots, a_N)^{\intercal} \in \mathbb{R}^N$ *or* $a \equiv (a_1, \ldots, a_N)^{\intercal} \in \mathbb{Q}^N$.

In words, the map $X : \Omega \to \mathbb{R}^N$ is an $\mathcal{E}$-random vector according to whether the events $\{a < X < b\}$, $\{a \leq X \leq b\}$, $\{a < X \leq b\}$, $\{a \leq X < b\}$ of Ω are in the σ-algebra $\mathcal{E}$, on varying of a, $b \in \mathbb{R}^N$ or a, $b \in \mathbb{Q}^N$. This means that the observer of the random phenomenon, in light of the information $\mathcal{E}$ available to her, can decide unambiguously whether the events $\{a < X < b\}$, $\{a \leq X \leq b\}$, $\{a < X \leq b\}$, $\{a \leq X < b\}$ occur or not, on varying of a, $b \in \mathbb{R}^N$ or, possibly, a, $b \in \mathbb{Q}^N$. An equivalent condition is that the observer, in light of her information $\mathcal{E}$, can decide unambiguously whether the events $\{X < a\}$, $\{X \leq a\}$, $\{X > a\}$, $\{X \geq a\}$ occur or not, on varying of $a \in \mathbb{R}^N$ or $\in \mathbb{Q}^N$.

Note that, in case we have $a_{K_0} > b_{K_0}$ for some $K_0 \in \{1, \ldots, N\}$, the intervals $\mathsf{X}_{K=1}^{N} (a_K, b_K)$, $\mathsf{X}_{K=1}^{N} [a_K, b_K]$, $\mathsf{X}_{K=1}^{N} (a_K, b_K]$, $\mathsf{X}_{K=1}^{N} [a_K, b_K)$ degenerate to the empty subset $\varnothing$ of $\mathbb{R}$ and the events referred to in Equations (6.15)-(6.18) become the impossible event $\varnothing$ of $\nleqslant$. In this case, Equations (6.15)-(6.18) hold trivially true.

As a consequence of Remarks 692, 693, and 694, we have

Proposition 697 *The map* $X : \Omega \to \mathbb{R}^N$ *is an* $\mathcal{E}$*-random vector if and only if the entries* $X_K : \Omega \to \mathbb{R}$ *of* X *are* $\mathcal{E}$*-random variables, for every* $K = 1, \ldots, N$.

Proof. Assume the map $X : \Omega \to \mathbb{R}^N$ is an $\mathcal{E}$-random vector. Fix any $K_0 \in \{1, \ldots, N\}$ and consider the entry $X_{K_0} : \Omega \to \mathbb{R}$ of X. We will have proved that X_{K_0} is an $\mathcal{E}$-random variable by showing that

$$\{a < X_{K_0} < b\} \in \mathcal{E} \tag{6.23}$$

for all $a, b \in \mathbb{R}$. Since $X : \Omega \to \mathbb{R}^N$ is an $\mathcal{E}$-random vector, we know that

$$\begin{aligned}
&\{-n < X_1 < n, \ldots, -n < X_{K_0-1} < n, a < X_{K_0} < b, -n < X_{K_0+1} < n, \ldots, -n < X_N < n\} \\
&\equiv \left\{X \in (\mathsf{X}_{K=1}^{K_0-1} (-n, n)) \times (a, b) \times \left(\mathsf{X}_{K=K_0+1}^{N} (-n, n)\right)\right\} \in \mathcal{E}
\end{aligned}$$

for every $n \in \mathbb{N}$ and all $a, b \in \mathbb{R}$. Hence,

$$\bigcup_{n=1}^{\infty} \left\{X \in (\mathsf{X}_{K=1}^{K_0-1} (-n, n)) \times (a, b) \times \left(\mathsf{X}_{K=K_0+1}^{N} (-n, n)\right)\right\} \in \mathcal{E} \tag{6.24}$$

for all $a, b \in \mathbb{R}$. On the other hand,

$$\begin{aligned}
&\bigcup_{n=1}^{\infty} \left\{X \in (\mathsf{X}_{K=1}^{K_0-1} (-n, n)) \times (a, b) \times \left(\mathsf{X}_{K=K_0+1}^{N} (-n, n)\right)\right\} \\
&= \left\{X \in \bigcup_{n=1}^{\infty} \left((\mathsf{X}_{K=1}^{K_0-1} (-n, n)) \times (a, b) \times \left(\mathsf{X}_{K=K_0+1}^{N} (-n, n)\right)\right)\right\} \\
&= \left\{X \in \left(\left(\mathsf{X}_{K=1}^{K_0-1} \bigcup_{n=1}^{\infty} (-n, n)\right) \times \bigcup_{n=1}^{\infty} (a, b) \times \left(\mathsf{X}_{K=K_0+1}^{N} \bigcup_{n=1}^{\infty} (-n, n)\right)\right)\right\} \\
&= \left\{X \in \left(\left(\mathsf{X}_{K=1}^{K_0-1} \mathbb{R}\right) \times (a, b) \times \left(\mathsf{X}_{K=K_0+1}^{N} \mathbb{R}\right)\right)\right\} \\
&= \{X_1 \in \mathbb{R}, \ldots, X_{K_0-1} \in \mathbb{R}, X_{K_0} \in (a, b), X_{K_0+1} \in \mathbb{R}, \ldots, X_N \in \mathbb{R}, \} \\
&= \bigcap_{K=1}^{K_0-1} \{X_K \in \mathbb{R}\} \cap \{X_{K_0} \in (a, b)\} \cap \bigcap_{K=K_0+1}^{N} \{X_K \in \mathbb{R}\} \\
&= \left(\bigcap_{K=1}^{K_0-1} \Omega\right) \cap \{X_{K_0} \in (a, b)\} \cap \left(\bigcap_{K=K_0+1}^{N} \Omega\right) \\
&= \{X_{K_0} \in (a, b)\} \equiv \{a < X_{K_0} < b\}.
\end{aligned} \tag{6.25}$$

Combining (6.24) and (6.25), it follows the desired (6.23).

Conversely, if $X_1, \ldots, X_N$ are $\mathcal{E}$-random variables, for every $K = 1, \ldots, N$, then we have

$$\{a_K < X_K < b_K\} \in \mathcal{E}$$

for every $K = 1, \ldots, N$, and for all $a_1, \ldots, a_N, b_1, \ldots, b_N \in \mathbb{R}$. It then follows

$$\bigcap_{K=1}^{N} \{a_K < X_K < b_K\} \in \mathcal{E}, \tag{6.26}$$

for all $a_1, \ldots, a_N, b_1, \ldots, b_N \in \mathbb{R}$. On the other hand,

$$\bigcap_{K=1}^{N} \{a_K < X_K < b_K\} = \{a_1 < X_1 < b_1, \ldots, a_N < X_N < b_N\} \equiv \{a < X < b\}, \tag{6.27}$$

where $a \equiv (a_1, \ldots, a_N)$ and $b \equiv (b_1, \ldots, b_N)$. Combining (6.26) and (6.27), we obtain that

$$\{a < X < b\} \in \mathcal{E},$$

for all $a \equiv (a_1, \ldots, a_N), b \equiv (b_1, \ldots, b_N) \in \mathbb{R}^N$. Thus, $X : \Omega \to \mathbb{R}^N$ is an $\mathcal{E}$-random vector. □

Corollary 698 *The map $X : \Omega \to \mathbb{R}^N$ is an $\mathcal{E}$-random vector if and only if the map $X_{K_1,\ldots,K_M} : \Omega \to \mathbb{R}^M$ with entries $X_{K_1,\ldots,K_M,1} \equiv \pi_{K_1}(X) \equiv X_{K_1}, \ldots, X_{K_1,\ldots,K_M,M} \equiv \pi_{K_M}(X) \equiv X_{K_M}$ is an $\mathcal{E}$-random vector on Ω with states in $\mathbb{R}^M$, for any $M \in \mathbb{N}$ and any choice of $K_1, \ldots, K_M \in \{1, \ldots, N\}$.*

Proof. . □

Let $\mathbb{R}^M$ [resp. $\mathbb{R}^N$] be the real M-dimensional [resp. N-dimensional] Euclidean space for some $M \in \mathbb{N}$ [resp. $M \in \mathbb{N}$], let $\mathcal{B}(\mathbb{R}^M)$ [resp. $\mathcal{B}(\mathbb{R}^N)$] the Borel family of $\mathbb{R}^M$ [resp. $\mathbb{R}^N$], and let $g : \mathbb{R}^M \to \mathbb{R}^N$ be a map from $\mathbb{R}^M$ to $\mathbb{R}^N$.

Definition 699 *We say that $g : \mathbb{R}^M \to \mathbb{R}^N$ is a* Borel map *if*

$$g^{-1}(B) \in \mathcal{B}(\mathbb{R}^M), \quad \forall B \in \mathcal{B}(\mathbb{R}^N). \tag{6.28}$$

Let $X : \Omega \to \mathbb{R}^M$ be an be an $\mathcal{E}$-random vector on Ω with states in $\mathbb{R}^M$ and let $g : \mathbb{R}^M \to \mathbb{R}^N$ be a Borel map from $\mathbb{R}^M$ to $\mathbb{R}^N$.

Proposition 700 *The map $g \circ X : \Omega \to \mathbb{R}^N$ given by*

$$(g \circ X)(\omega) \overset{def}{=} g(X(\omega)), \quad \forall \omega \in \Omega, \tag{6.29}$$

is an $\mathcal{E}$-random vector on Ω with states in $\mathbb{R}^N$.

Proof. . □

Remark 701 *We have*

$$g(X(\omega)) \equiv (g_1(X(\omega)), \ldots, g_N(X(\omega))) \equiv (g_1(X_1(\omega), \ldots, X_M(\omega)), \ldots, g_N(X_1(\omega), \ldots, X_M(\omega)))$$

for every $\omega \in \Omega$, where $g_1, \ldots, g_N$ are the entries of g and $X_1, \ldots, X_M$ are the entries of X.

6.2 Joint and Marginal Distribution, Distribution Function, and Density

Let $X:\Omega\to\mathbb{R}^N$ be an $\mathcal{E}$-random vector with entries $X_1,\ldots,X_N$. For any $M\in\mathbb{N}$ and any choice of $K_1,\ldots,K_M\in\{1,\ldots,N\}$ write $X_{K_1,\ldots,K_M}:\Omega\to\mathbb{R}^M$ for the $\mathcal{E}$-random vector on Ω with states in $\mathbb{R}^M$ and entries $X_{K_1,\ldots,K_M,1}\equiv\pi_{K_1}(X)\equiv X_{K_1},\ldots,X_{K_1,\ldots,K_M,M}\equiv\pi_{K_M}(X)\equiv X_{K_M}$, that is $X_{K_1,\ldots,K_M}\equiv(X_{K_1},\ldots,X_{K_M})^{\intercal}$.

Definition 702 *The distribution $P_X:\mathcal{B}(\mathbb{R}^N)\to\mathbb{R}_+$ of X, briefly P_X, given by*

$$P_X(B)\overset{def}{=}\mathbf{P}(X\in B),\quad\forall B\in\mathcal{B}(\mathbb{R}^N),$$

is also called the joint distribution *or* joint law *of $X_1,\ldots,X_N$ and also denoted by $P_{X_1,\ldots,X_N}:\mathcal{B}(\mathbb{R}^N)\to\mathbb{R}_+$, briefly $P_{X_1,\ldots,X_N}$. More generally, we call the* joint distribution *or* joint law *of $X_{K_1},\ldots,X_{K_M}$ the distribution $P_{X_{K_1,\ldots,K_M}}:\mathcal{B}(\mathbb{R}^M)\to\mathbb{R}_+$, briefly $P_{X_{K_1,\ldots,K_M}}$, of the $\mathcal{E}$-random vector $X_{K_1,\ldots,K_M}:\Omega\to\mathbb{R}^M$, given by*

$$P_{X_{K_1,\ldots,K_M}}(B)\overset{def}{=}\mathbf{P}(X_{K_1,\ldots,K_M}\in B),\quad\forall B\in\mathcal{B}(\mathbb{R}^M).$$

The distribution $P_{X_{K_1,\ldots,K_M}}:\mathcal{B}(\mathbb{R}^M)\to\mathbb{R}_+$ is also denoted by $P_{X_{K_1},\ldots,X_{K_M}}:\mathcal{B}(\mathbb{R}^M)\to\mathbb{R}_+$, briefly $P_{X_1,\ldots,X_N}$. In case $M<N$ and $K_1<\cdots<K_M$, the distribution $P_{X_{K_1,\ldots,K_M}}$ is also called the marginal distribution *or* marginal law *of X referred to the indices $K_1,\ldots,K_M$. In particular, for every $K=1,\ldots,N$, the marginal distribution of X referred to the index K is just the distribution $P_{X_K}:\mathcal{B}(\mathbb{R})\to\mathbb{R}_+$, briefly P_{X_K}, of the entry X_K of X and it is also called the Kth* marginal distribution *of X.*

Remark 703 *Assume the $\mathcal{E}$-random vector $X\equiv(X_1,\ldots,X_N)^{\intercal}$ is discrete. Then, we can write $X(\Omega)\equiv(x_j)_{j\in\mathbb{J}}$ for some$\mathbb{J}\subseteq\mathbb{N}$, where $x_j\in\mathbb{R}^N$, for every $j\in\mathbb{J}$. In this case, setting $E_j\equiv\{X=x_j\}$, we have $E_j\in\mathcal{E}$, for every $j\in\mathbb{J}$, and*

$$P_X(B)=\textstyle\sum_{j\in\mathbb{J}:x_j\in B}P_X(X=x_j),$$

for every $B\in\mathbb{R}^N$.

Definition 704 *We call the* distribution function *of X the function $F_X:\mathbb{R}^N\to\mathbb{R}_+$, briefly F_X, given by*

$$F_X(x_1,\ldots,x_N)\overset{def}{=}P_X\left(\mathsf{X}_{K=1}^{N}(-\infty,x_K]\right),\quad\forall(x_1,\ldots,x_N)\in\mathbb{R}^N.$$

The distribution function F_X is also called the joint distribution function *of $X_1,\ldots,X_N$ also denoted by $F_{X_1,\ldots,X_N}:\mathbb{R}^N\to\mathbb{R}_+$, briefly $F_{X_1,\ldots,X_N}$. More generally, we call the* joint distribution function *of $X_{K_1},\ldots,X_{K_M}$ the distribution function $F_{X_{K_1,\ldots,K_M}}:\mathbb{R}^M\to\mathbb{R}_+$, briefly $F_{X_{K_1,\ldots,K_M}}$, of the $\mathcal{E}$-random vector $X_{K_1,\ldots,K_M}:\Omega\to\mathbb{R}^M$ given by*

$$F_{X_{K_1,\ldots,K_M}}(x_1,\ldots,x_M)=P_{X_{K_1,\ldots,K_M}}\left(\mathsf{X}_{J=1}^{M}(-\infty,x_J]\right),\quad\forall(x_1,\ldots,x_M)\in\mathbb{R}^M.$$

The distribution function $F_{X_{K_1,\ldots,K_M}}:\mathbb{R}^M\to\mathbb{R}$ is also denoted by $F_{X_{K_1},\ldots,X_{K_M}}:\mathbb{R}^M\to\mathbb{R}_+$, briefly $F_{X_{K_1},\ldots,X_{K_M}}$. In case $M<N$ and $K_1<\cdots<K_M$, the distribution function $F_{X_{K_1,\ldots,K_M}}$ is also called the marginal distribution function *of X referred to the indices $K_1,\ldots,K_M$. In particular, for every $K=1,\ldots,N$, the marginal distribution function of X referred to the index K is just the distribution function $F_{X_K}:\mathbb{R}\to\mathbb{R}$, briefly F_{X_K}, of the entry X_K of X and it is also called the Kth* marginal distribution function *of X.*

Remark 705 *We clearly have*

$$F_X(x_1,\ldots,x_N)=\mathbf{P}(X_1\leq x_1,\ldots,X_N\leq x_N)=\mathbf{P}\left(\bigcap_{K=1}^{N}\{X_K\leq x_K\}\right)$$

for every $(x_1,\ldots,x_N)\in\mathbb{R}^N$ *and*

$$F_{X_{K_1},\ldots,K_M}(x_1,\ldots,x_M)=\mathbf{P}(X_{K_1}\leq x_1,\ldots,X_{K_M}\leq x_M)=\mathbf{P}\left(\bigcap_{J=1}^{M}\{X_{K_J}\leq x_J\}\right)$$

for every $(x_1,\ldots,x_M)\in\mathbb{R}^M$.

Let $F_{X_1,\ldots,X_N}:\mathbb{R}^N\to\mathbb{R}_+$ [resp. $F_{X_{K_1},\ldots,X_{K_M}}:\mathbb{R}^M\to\mathbb{R}_+$] be the joint distribution function of $X_1,\ldots,X_N$ [res. of $X_{K_1},\ldots,X_{K_M}$, for a fixed $M\in\mathbb{N}$ and a choice of $K_1,\ldots,K_M\in\{1,\ldots,N\}$].

Proposition 706 *We have*

$$\lim_{x_1\to+\infty}\ldots\lim_{x_N\to+\infty}F_{X_1,\ldots,X_N}(x_1,\ldots,x_N)=1$$

and

$$\lim_{x_K\to-\infty}F_{X_1,\ldots,X_K,\ldots,X_N}(x_1,\ldots,x_K,\ldots,x_N)=0,$$

for any fixed $K=1,\ldots,N$. *Similarly,*

$$\lim_{x_1\to+\infty}\ldots\lim_{x_M\to+\infty}F_{X_{K_1},\ldots,X_{K_M}}(x_1,\ldots,x_M)=1$$

and

$$\lim_{x_J\to-\infty}F_{X_{K_1},\ldots,X_{K_J},\ldots,X_M}(x_1,\ldots,x_J,\ldots,x_M)=0,$$

for any fixed $J=1,\ldots,M$.

Proof. . □

Proposition 707 *The distribution function* $F_{X_1,\ldots,X_N}:\mathbb{R}^N\to\mathbb{R}_+$ *is a non-decreasing function of each of its arguments. In symbols,*

$$F_{X_1,\ldots,X_K,\ldots,X_N}(x_1,\ldots,x'_K,\ldots,x_N)\leq F_{X_1,\ldots,X_K,\ldots,X_N}(x_1,\ldots,x''_K,\ldots,x_N),$$

for every fixed $K=1,\ldots,N$, *all* $x'_K,x''_K\in\mathbb{R}$ *such that* $x'_K\leq x''_K$ *and every* $x_J\in\mathbb{R}$, *on varying of* $J=1,\ldots,N$, *such that* $J\neq K$. *The same result holds true for the distribution function* $F_{X_{K_1},\ldots,X_{K_M}}:\mathbb{R}^M\to\mathbb{R}_+$.

Proposition 708 *Fixed any* $K=1,\ldots,N$ *we have*

$$\lim_{x_1\to+\infty}\ldots\lim_{x_{K-1}\to+\infty}\lim_{x_{K+1}\to+\infty}\ldots\lim_{x_N\to+\infty}F_{X_1,\ldots,X_{K-1},X_K,X_{K+1},\ldots,X_N}(x_1,\ldots,x_{K-1},x_K,x_{K+1},\ldots,x_N)$$
$$=F_{X_K}(x_K),$$

where the limits are considered for all variables $x_1, \ldots, x_N$ *except* x_K *and* F_{X_K} *is the* K*th* marginal distribution function *of* X*. More generally, we have*

$$\lim_{x_1 \to +\infty} \ldots \lim_{x_{K_1-1} \to +\infty} \lim_{x_{K_1+1} \to +\infty} \ldots \lim_{x_{K_M-1} \to +\infty} \lim_{x_{K_M+1} \to +\infty} \ldots \lim_{x_N \to +\infty}$$
$$F_{X_1,\ldots,X_{K_1-1},X_{K_1},X_{K_1+1},\ldots,X_{K_M-1},X_{K_M},X_{K_M+1},\ldots,X_N}(x_1, \ldots, x_{K_1-1}, x_{K_1}, x_{K_1+1}, \ldots, x_{K_M-1}, x_{K_M}, x_{K_M+1}, \ldots, x_N)$$
$$= F_{X_{K_1},\ldots,X_{K_M}}(x_{K_1}, \ldots, x_{K_M}),$$

where the limits are considered for all variables $x_1, \ldots, x_N$ *except* $x_{K_1}, \ldots, x_{K_M}$ *and* $F_{X_{K_1},\ldots,X_{K_M}}$ *is the* marginal distribution function *of* X *referred to the indices* $K_1, \ldots, K_M$*.*

Proof. . □

Let $X : \Omega \to \mathbb{R}^N$ be an $\mathcal{E}$-random vector and let $P_X : \mathcal{B}(\mathbb{R}^N) \to \mathbb{R}_+$ [resp. $F_X : \mathbb{R}^N \to \mathbb{R}_+$] be the distribution [resp. the distribution function] of X.

Definition 709 *We say that* $P_X : \mathcal{B}(\mathbb{R}^N) \to \mathbb{R}_+$ *is* absolutely continuous *with respect to the Borel-Lebesgue measure* $\mu_L^N : \mathcal{B}(\mathbb{R}^N) \to \mathbb{R}_+$*, if we have*

$$P_X(B) = 0, \quad \forall B \in \mathcal{B}(\mathbb{R}^N) : \mu_L^N(B) = 0.$$

Theorem 710 (Radon-Nikodým) *The distribution* $P_X : \mathcal{B}(\mathbb{R}^N) \to \mathbb{R}_+$ *is absolutely continuous with respect to the Lebesgue measure* $\mu_L^N : \mathcal{B}(\mathbb{R}^N) \to \mathbb{R}_+$ *on* $\mathbb{R}^N$ *if and only if there exists a Lebesgue-integrable function* $f : \mathbb{R}^N \to \mathbb{R}_+$*, briefly* f*, such that*

$$P_X(B) = \int_B f(x_1, \ldots, x_N)\, d\mu_L^N(x_1, \ldots, x_N), \tag{6.30}$$

for every $B \in \mathcal{B}(\mathbb{R}^N)$*. Such a function* f *is the Radon-Nykodym derivative of* P_X *with respect to* μ_L^N *and it is uniquely determined almost everywhere; namely, if* $\tilde{f} : \mathbb{R}^N \to \mathbb{R}_+$ *is another Lebesgue-integrable real function on* $\mathbb{R}^N$ *satisfying (6.30) we have*

$$f = \tilde{f},$$

μ_L^N*-almost everywhere on* $\mathbb{R}^N$*, briefly* μ_L^N*-a.e. on* $\mathbb{R}^N$*, which meas that* $\mu_L^N\left(\left\{x \in \mathbb{R}^N : f(x) \neq \tilde{f}(x)\right\}\right) = 0$*.*

Definition 711 *We say that the random vector* $X : \Omega \to \mathbb{R}^N$ *is* absolutely continuous *if the distribution* $P_X : \mathcal{B}(\mathbb{R}^N) \to \mathbb{R}_+$ *of* X *is absolutely continuous with respect to the Borel-Lebesgue measure* $\mu_L^N : \mathcal{B}(\mathbb{R}^N) \to \mathbb{R}_+$*.*

Definition 712 *If the random vector* $X : \Omega \to \mathbb{R}^N$ *is absolutely continuous, we call a* version of the density *of* X *any Radon-Nykodym derivative of* P_X *with respect to* μ_L^N*. We call the* density *of* X *the equivalence class of all Lebesgue integrable real functions on* $\mathbb{R}^N$ *satisfying (6.30) which are almost everywhere equal. In measure theory, such an equivalence class is denoted by* $\frac{dP_X}{d\mu_{L^N}}$*. A version of the density [resp. the density] of* X *is also called a* version of the joint density *[resp. the* joint density*] of the entries* $X_1, \ldots, X_N$ *of* X*.*

However, likewise the one-dimensional case, for our purposes it is possible to neglect the distinction between the density of a random variable and its versions. Therefore, we will refer to any version of the density of the random vector X as to the *density* of X, and we will denote it by the symbol $f_X : \mathbb{R}^N \to \mathbb{R}$.

Let $X : \Omega \to \mathbb{R}^N$ be an absolutely continuous $\mathcal{E}$-random vector and let $f_X : \mathbb{R}^N \to \mathbb{R}$ be the density of X.

Remark 713 *We have*

$$f_X(x_1, \ldots, x_N) \geq 0, \quad \mu_L^N\text{-a.e. on } \mathbb{R}^N, \tag{6.31}$$

and

$$\int_{\mathbb{R}^N} f_X(x_1, \ldots, x_N)\, d\mu_L^N(x_1, \ldots, x_N) = 1. \tag{6.32}$$

Proposition 714 *If the $\mathcal{E}$-random vector $X : \Omega \to \mathbb{R}^N$ on Ω with states in $\mathbb{R}^N$ is absolutely continuous with density $f_X : \mathbb{R}^N \to \mathbb{R}$, Then, all the entries $X_1, \ldots, X_N$ of X are absolutely continuous. Furthermore, fixed any $K = 1, \ldots, N$, the density $f_{X_K} : \mathbb{R} \to \mathbb{R}$ of the Kth entry X_K of X is given by*

$$f_{X_K}(x_K) \overset{def}{=} \int_{\mathbb{R}^{N-1}} f_X(x_1, \ldots, x_K, \ldots, x_N)\, d\mu_{L^{N-1}}(x_1, \ldots, \mathring{x}_K, \ldots, x_N) \tag{6.33}$$

fro every $x_K \in \mathbb{R}$, where the symbol $\mathring{x}_K$ denotes that x_K does not appear among the integration variables. More generally, the $\mathcal{E}$-random vector $X_{K_1,\ldots,K_M} : \Omega \to \mathbb{R}^M$ is absolutely continuous and the density $f_{X_{K_1,\ldots,K_M}} : \mathbb{R}^M \to \mathbb{R}_+$ of $X_{K_1,\ldots,K_M}$, is given by

$$f_{X_{K_1,\ldots,K_M}}(x_{K_1}, \ldots, x_{K_M}) \overset{def}{=} \int_{\mathbb{R}^{N-M}} f_X(x_1, \ldots, x_{K_1}, \ldots, x_{K_M}, \ldots, x_N)\, d\mu_L^{N-M}(x_1, \ldots, \mathring{x}_{K_1}, \ldots, \mathring{x}_{K_M}, \ldots, x_N)$$

for every $x_{K_1}, \ldots, x_{K_M} \in \mathbb{R}^M$, where the symbols $\mathring{x}_{K_1}, \ldots, \mathring{x}_{K_M}$ denote that $x_{K_1}, \ldots, x_{K_M}$ do not appear among the integration variables.

Proof. . □

Definition 715 *Fixed any $K = 1, \ldots, N$, the density $f_{X_K} : \mathbb{R} \to \mathbb{R}$ of the Kth entry X_K of X, briefly f_{X_K}, is also called the Kth* marginal density *of X.*

Definition 716 *In case $M < N$ and $K_1 < \cdots < K_M$, the density of the random vector $X_{K_1,\ldots,K_M} \equiv (X_{K_1}, \ldots, X_{K_M})^\intercal$ is also called the* marginal density *of X referred to the indices $K_1, \ldots, K_M$, or the* joint density of the entries $X_{K_1}, \ldots, X_{K_M}$ *of X and also denoted by $f_{X_{K_1},\ldots,X_{K_M}} : \mathbb{R}^M \to \mathbb{R}$, briefly $f_{X_{K_1},\ldots,X_{K_M}}$.*

Proposition 717 *The $\mathcal{E}$-random vector $X : \Omega \to \mathbb{R}^N$, with distribution function $F_X : \mathbb{R}^N \to \mathbb{R}$, is absolutely continuous, with density function $f_X : \mathbb{R}^N \to \mathbb{R}$, if and only if we have*

$$\frac{\partial^N F_X}{\partial x_1 \ldots \partial x_N}(x_1, \ldots, x_N) = f_X(x_1, \ldots, x_N),$$

μ_L^N-a.e. on $\mathbb{R}^N$, and

$$F_X(x_1,\ldots,x_N)=\int_{\mathsf{X}_{N=1}^{N}(-\infty,x_N]} f_X(u_1,\ldots,u_N)\,d\mu_L^N(u_1,\ldots,u_N),$$

for every $(x_1,\ldots,x_N)\in\mathbb{R}^N$. More generally, the $\mathcal{E}$-random vector $X_{K_1,\ldots,K_M}:\Omega\to\mathbb{R}^M$, where $X_{K_1,\ldots,K_M}\equiv\left(X_{K_1},\ldots,X_{K_M}\right)$, with distribution function $F_{X_{K_1},\ldots,K_M}:\mathbb{R}^M\to\mathbb{R}_+$, is absolutely continuous, with density function $f_{X_{K_1},\ldots,X_{K_M}}:\mathbb{R}^M\to\mathbb{R}$, if and only if we have

$$\frac{\partial^M F_{X_{K_1},\ldots,K_M}}{\partial x_{K_1}\ldots\partial x_{K_M}}(x_{K_1},\ldots,x_{K_M})=f_{X_{K_1},\ldots,X_{K_M}}(x_{K_1},\ldots,x_{K_M}),$$

μ_L^M-a.e. on $\mathbb{R}^M$, and

$$F_{X_{K_1},\ldots,K_M}(x_{K_1},\ldots,x_{K_M})=\int_{\mathsf{X}_{J=1}^{M}\left(-\infty,x_{K_J}\right]} f_{X_{K_1},\ldots,X_{K_M}}(u_{K_1},\ldots,u_{K_M})\,d\mu_L^N(u_{K_1},\ldots,u_{K_M}),$$

for every $(x_{K_1},\ldots,x_{K_M})\in\mathbb{R}^M$.

Proof. . □

As a particular case of Proposition 717, if the random vector $X\equiv(X_1,\ldots,X_N)^\intercal$ is absolutely continuous, Then, all the entries are. In general, the converse is not true.

Example 718 *Let $Z:\Omega\to\mathbb{R}$ be an absolutely continuous random variable with density $f_Z:\mathbb{R}\to\mathbb{R}$. Consider the random vector $X\equiv(X_1,X_2)^\intercal$, where $X_k=Z$, for $k=1,2$. Then, X is not absolutely continuous.*

Discussion. Considering the subset D of $\mathbb{R}^2$ given by

$$D\overset{\text{def}}{=}\left\{(x_1,x_2)\in\mathbb{R}^2:x_1=x_2\right\},$$

we have that $D\in\mathcal{B}(\mathbb{R}^2)$ and

$$\{X\in D\}=\{X_1=X_2\}.$$

In fact, $\omega\in\{X\in D\}$ if and only if $X(\omega)\equiv(X_1(\omega),X_2(\omega))\in D$, that is, if and only if $X_1(\omega)=X_2(\omega)$, that is, if and only if $\omega\in\{X_1=X_2\}$. Now,

$$\mathbf{P}(X_1=X_2)=\mathbf{P}(Z=Z)=1.$$

Hence, considering the distribution $P_X:\mathcal{B}(\mathbb{R}^2)\to\mathbb{R}$ of the random vector X, we have

$$P_X(D)=\mathbf{P}(X_1=X_2)=1.$$

On the other hand,

$$\mu_L^2(D)=0.$$

It follows that $P_X:\mathcal{B}(\mathbb{R}^2)\to\mathbb{R}$ is not absolutely continuous with respect to $\mu_L^2:\mathcal{B}(\mathbb{R}^2)\to\mathbb{R}$.

Under the hypothesis that $f_Z : \mathbb{R} \to \mathbb{R}$ is continuous, for instance $Z \tilde{} N(0,1)$, it is possible to show the lack of absolute continuity of the random vector X by a different argument. In fact, consider the distribution function $F_Z : \mathbb{R} \to \mathbb{R}$ of the random vector Z, we have

$$\begin{aligned} F_Z(x_1, x_2) &= \mathbf{P}(X_1 \leq x_1, X_2 \leq x_2) = \mathbf{P}(Z \leq x_1, Z \leq x_2) \\ &= \mathbf{P}(Z \leq x_1 \wedge x_2) = F_Z(x_1 \wedge x_2) \\ &= \int_{(-\infty, x_1 \wedge x_2]} f_Z(z)\, d\mu_L(z) = \int_{-\infty}^{x_1 \wedge x_2} f_Z(z)\, dz, \end{aligned}$$

for all $x_1, x_2 \in \mathbb{R}$. It follows,

$$F_X(x_1, x_2) = \begin{cases} \int_{-\infty}^{x_1} f_Z(z)\, dz, & \text{if } x_1 \leq x_2, \\ \int_{-\infty}^{x_2} f_Z(z)\, dz, , & \text{if } x_1 > x_2. \end{cases}$$

Therefore,

$$\left(\frac{\partial F_X}{\partial x_2}\right)_{(x_1,x_2)} = \begin{cases} 0, & \text{if } x_1 < x_2, \\ f_Z(x_2), & \text{if } x_1 > x_2. \end{cases} \quad \text{and} \quad \left(\frac{\partial F_X}{\partial x_1}\right)_{(x_1,x_2)} = \begin{cases} f_Z(x_1), & \text{if } x_1 < x_2, \\ 0, & \text{if } x_1 > x_2. \end{cases}$$

As a consequence,

$$\left(\frac{\partial F_X}{\partial x_1 \partial x_2}\right)_{(x_1,x_2)} = 0,$$

for every $(x_1, x_2) \in \mathbb{R}^2 - \{(x_1, x_2) : x_1 = x_2\}$. On the other hand,

$$\mu_L^2(D) = 0.$$

Therefore, if Z were absolutely continuous the function $f_X : \mathbb{R} \to \mathbb{R}$ given by

$$f_X(x_1, x_2) \overset{\text{def}}{=} 0, \quad \forall (x_1, x_2) \in \mathbb{R}^2$$

would be the density function of X. We would then obtain the equalities

$$f_{X_1}(x_1) = \int_{\mathbb{R}} f_X(x_1, x_2)\, d\mu_L(x_2) = 0 \quad \text{and} \quad f_{X_2}(x_2) = \int_{\mathbb{R}} f_X(x_1, x_2)\, d\mu_L(x_1) = 0.$$

for all $x_1, x_2 \in \mathbb{R}$, which are clearly absurd. □

Let $X : \Omega \to \mathbb{R}^M$ be an $\mathcal{E}$-random vector and let $g : \mathbb{R}^M \to \mathbb{R}^N$ be a Borel map with entries $g_1, \ldots, g$ (see Remark 701).

Proposition 719 *The $\mathcal{E}$-random vector $g \circ X : \Omega \to \mathbb{R}^N$ is Lebesgue integrable on Ω (with respect to $\mathbf{P}$), that is*

$$\int_\Omega |g_K \circ X|\, d\mathbf{P} < \infty, \tag{6.34}$$

for every $K = 1, \ldots, N$, if and only if g is Lebesgue integrable on $\mathbb{R}^M$ with respect to P_X, that is

$$\int_{\mathbb{R}^N} |g_K(x_1, \ldots, x_M)|\, dP_X < \infty, \tag{6.35}$$

for every $K = 1, \ldots, N$. In this case, we have

$$\int_\Omega g_K \circ X d\mathbf{P} = \int_{\mathbb{R}^M} g_K(x_1, \ldots, x_M)\, dP_X \tag{6.36}$$

for every $K = 1, \ldots, N$.

Proposition 720 *Assume $X : \Omega \to \mathbb{R}^M$ is absolutely continuous with density $f_X : \mathbb{R}^M \to \mathbb{R}$. Then, the $\mathcal{E}$-random vector $g \circ X : \Omega \to \mathbb{R}^N$ is Lebesgue integrable on Ω (with respect to $\mathbf{P}$) if and only if the product $|g_K|\ f$ is Lebesgue integrable on Ω with respect to the Lebesgue measure, for every $K = 1, \ldots, N$, that is*

$$\int_{\mathbb{R}^M} |g_K(x_1, \ldots, x_M)| f_X(x_1, \ldots, x_M) d\mu_L^M(x_1, \ldots, x_M) < \infty,$$

for every $K = 1, \ldots, N$. In this case, we have

$$\int_\Omega g_K \circ X d\mathbf{P} = \int_{\mathbb{R}^M} g_K(x_1, \ldots, x_M) f_X(x_1, \ldots, x_M) d\mu_L^M(x_1, \ldots, x_M).$$

Let $f : \mathbb{R}^N \to \mathbb{R}$ be a real function on $\mathbb{R}^N$.

Definition 721 *We say that f is an N-*dimensional density*, if f satisfies Equations (6.31) and (6.31).*

Theorem 722 *For any N-dimensional density f there exist a probability space $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ and an absolutely continuous $\mathcal{E}$-random vector $X : \Omega \to \mathbb{R}^N$ with density $f_X : \mathbb{R}^N \to \mathbb{R}_+$ such that*

$$f_X = f,$$

μ_L^N-a.e. on $\mathbb{R}^N$.

Problem 723 *Let $X : \Omega \to \mathbb{R}^2$ be an absolutely continuous random vector with density $f_X : \mathbb{R}^2 \to \mathbb{R}_+$ given by*

$$f_X(x_1, x_2) \overset{def}{=} \begin{cases} \frac{1}{2} & \text{if } (x_1, x_2) \in [0,1] \times [0,2] \\ 0 & \text{otherwise} \end{cases}.$$

Compute the distribution function $F_X : \mathbb{R}^2 \to \mathbb{R}_+$ of X. Compute also the marginal densities and distribution functions of the entries X_1 and X_2 of X. In the end, compute the probabilities $\mathbf{P}(-1 \leq X_1 \leq 2, -1 \leq X_2 \leq 1)$, $\mathbf{P}(-1 \leq X_1 \leq 2)$, and $\mathbf{P}(-1 \leq X_1 \leq 1)$ and check whether the equality

$$\mathbf{P}(-1 \leq X_1 \leq 2, -1 \leq X_2 \leq 1) = \mathbf{P}(-1 \leq X_1 \leq 2)\mathbf{P}(-1 \leq X_2 \leq 1)$$

holds true.

Solution. We have

$$F_X(x_1, x_2) \overset{\text{def}}{=} \int_{(-\infty, x_1] \times (-\infty, x_2]} f_X(u_1, u_2) d\mu_L^2(u_1, u_2), \quad \forall (x_1, x_2) \in \mathbb{R}^2.$$

On the other hand, since

$$f_X(x_1, x_2) = \frac{1}{2} 1_{[0,1] \times [0,2]}(x_1, x_2), \quad \forall (x_1, x_2) \in \mathbb{R}^2.$$

we can write

$$\begin{aligned}
&\int_{(-\infty,x_1]\times(-\infty,x_2]} f_X(u_1,u_2)\, d\mu_{L^2}(u_1,u_2)\\
&= \iint\limits_{(-\infty,x_1]\times(-\infty,x_2]} \frac{1}{2} 1_{[0,1]\times[0,2]}(u_1,u_2)\, d\mu_{L^2}(u_1,u_2)\\
&= \iint\limits_{\mathbb{R}^2} \frac{1}{2} 1_{[0,1]\times[0,2]}(u_1,u_2)\, 1_{(-\infty,x_1]\times(-\infty,x_2]}(u_1,u_2)\; d\mu_{L^2}(u_1,u_2)\\
&= \frac{1}{2} \iint\limits_{\mathbb{R}^2} 1_{((-\infty,x_1]\times(-\infty,x_2])\cap([0,1]\times[0,2])}(u_1,u_2)\, d\mu_{L^2}(u_1,u_2)\\
&= \frac{1}{2} \iint\limits_{\mathbb{R}^2} 1_{((-\infty,x_1]\cap[0,1])\times((-\infty,x_2]\cap[0,2])}(u_1,u_2)\, d\mu_{L^2}(u_1,u_2)\\
&= \frac{1}{2} \mu_{L^2}\left(((-\infty,x_1]\cap[0,1])\times((-\infty,x_2]\cap[0,2])\right).
\end{aligned}$$

Hence, we need to consider some cases depending on the position of the point (x_1,x_2) with respect to the set $[0,1]\times[0,2]$. In fact, we have

$$((-\infty,x_1]\cap[0,1])\times((-\infty,x_2]\cap[0,2]) = \begin{cases} \varnothing, & \text{if } x_1<0 \ \vee\ x_2<0,\\ [0,x_1]\times[0,x_2], & \text{if } 0\le x_1\le 1 \ \wedge\ 0\le x_2\le 2,\\ [0,x_1]\times[0,2], & \text{if } 0\le x_1\le 1 \ \wedge\ 2<x_2,\\ [0,1]\times[0,x_2], & \text{if } 1<x_1 \ \wedge\ 0\le x_2\le 2,\\ [0,1]\times[0,2], & \text{if } 1<x_1 \ \wedge\ 2<x_2. \end{cases}$$

Therefore,

$$\frac{1}{2}\mu_{L^2}\left(((-\infty,x_1]\cap[0,1])\times((-\infty,x_2]\cap[0,2])\right) = \begin{cases} \mu_{L^2}(\varnothing)=0, & \text{if } x_1<0 \ \vee\ x_2<0,\\ \mu_{L^2}([0,x_1]\times[0,x_2])=x_1x_2, & \text{if } 0\le x_1\le 1 \ \wedge\ 0\le x_2\\ \mu_{L^2}([0,x_1]\times[0,2])=2x_1, & \text{if } 0\le x_1\le 1 \ \wedge\ 2<x_2,\\ \mu_{L^2}([0,1]\times[0,x_2])=x_2, & \text{if } 1<x_1 \ \wedge\ 0\le x_2\le 2,\\ \mu_{L^2}([0,1]\times[0,2])=2, & \text{if } 1<x_1 \ \wedge\ 2<x_2. \end{cases}$$

It follows,

$$F_X(x_1,x_2) = \begin{cases} 0, & \text{if } x_1<0 \ \vee\ x_2<0,\\ \frac{1}{2}x_1x_2, & \text{if } 0\le x_1\le 1 \ \wedge\ 0\le x_2\le 2,\\ x_1, & \text{if } 0\le x_1\le 1 \ \wedge\ 2<x_2,\\ \frac{1}{2}x_2, & \text{if } 1<x_1 \ \wedge\ 0\le x_2\le 2,\\ 1, & \text{if } 1<x_1 \ \wedge\ 2<x_2. \end{cases}$$

Note that

$$\frac{\partial^2 F_X}{\partial x_1\partial x_2}(x_1,x_2) = \begin{cases} 0, & \text{if } x_1<0 \ \vee\ x_2<0,\\ \frac{1}{2}, & \text{if } 0<x_1<1 \ \wedge\ 0<x_2<2,\\ 0, & \text{if } 0<x_1<1 \ \wedge\ 2<x_2,\\ 0, & \text{if } 1<x_1 \ \wedge\ 0<x_2<2,\\ 0, & \text{if } 1<x_1 \ \wedge\ 2<x_2. \end{cases}$$

That is

$$\frac{\partial^2 F_X}{\partial x_1\partial x_2}(x_1,x_2) = f_X(x_1,x_2),$$

for every $(x_1, x_2) \in \mathbb{R}^2 - \partial([0,1] \times [0,2])$,where $\partial([0,1] \times [0,2])$ denotes the boundary of $[0,1] \times [0,2]$. Hence,

$$\frac{\partial^2 F_X}{\partial x_1 \partial x_2}(x_1, x_2) = f_X(x_1, x_2),$$

μ_L^2-a.e. on $\mathbb{R}^2$. Now, we have

$$f_{X_1}(x_1) = \int_{\mathbb{R}} f_X(x_1, x_2)\, d\mu_{L^1}(x_2) \quad \text{and} \quad f_{X_2}(x_2) = \int_{\mathbb{R}} f_X(x_1, x_2)\, d\mu_{L^1}(x_1),$$

for every $x_1 \in \mathbb{R}$ and $x_2 \in \mathbb{R}$, respectively, where

$$\begin{aligned}\int_{\mathbb{R}} f_X(x_1, x_2)\, d\mu_{L^1}(x_2) &= \frac{1}{2}\int_{\mathbb{R}} 1_{[0,1]\times[0,2]}(x_1, x_2)\; d\mu_{L^1}(x_2) = \frac{1}{2}\int_{\mathbb{R}} 1_{[0,1]}(x_1)\, 1_{[0,2]}(x_2)\; d\mu_{L^1}(x_2) \\ &= \frac{1}{2} 1_{[0,1]}(x_1) \int_{\mathbb{R}} 1_{[0,2]}(x_2)\; d\mu_{L^1}(x_2) = \frac{1}{2} 1_{[0,1]}(x_1)\, \mu_{L^1}([0,2]) = 1_{[0,1]}(x_1).\end{aligned}$$

and, similarly,

$$\int_{\mathbb{R}} f_X(x_1, x_2)\; d\mu_{L^1}(x_2) = \frac{1}{2} 1_{[0,2]}(x_2).$$

Hence,

$$f_{X_1}(x_1) = 1_{[0,1]}(x_1) \quad \text{and} \quad f_{X_2}(x_2) = \frac{1}{2} 1_{[0,2]}(x_2),$$

for every $x_1 \in \mathbb{R}$ and $x_2 \in \mathbb{R}$, respectively. In the end,

$$\begin{aligned}\mathbf{P}(-1 \le X_1 \le 2, -1 \le X_2 \le 1) &= \mathbf{P}(X \in [-1,2] \times [-1,1]) \\ &= \iint_{[-1,2]\times[-1,3]} \frac{1}{2} 1_{[0,1]\times[0,2]}(x_1, x_2)\, d\mu_{L^2}(x_1, x_2) \\ &= \frac{1}{2}\iint_{\mathbb{R}^2} 1_{[-1,2]\times[-1,1]}(x_1, x_2)\, 1_{[0,1]\times[0,2]}(x_1, x_2)\, d\mu_{L^2}(x_1, x_2) \\ &= \frac{1}{2}\iint_{\mathbb{R}^2} 1_{([-1,2]\times[-1,1])\cap([0,1]\times[0,2])}(x_1, x_2)\, d\mu_{L^2}(x_1, x_2) \\ &= \frac{1}{2}\iint_{\mathbb{R}^2} 1_{([-1,2]\cap[0,1])\times([-1,1]\cap[0,2])}(x_1, x_2)\, d\mu_{L^2}(x_1, x_2) \\ &= \frac{1}{2}\iint_{\mathbb{R}^2} 1_{[0,1]\times[0,1]}(x_1, x_2)\, d\mu_{L^2}(x_1, x_2) \\ &= \frac{1}{2}\mu_{L^2}([0,1] \times [0,1]) \\ &= \frac{1}{2}\end{aligned}$$

while

$$\begin{aligned}\mathbf{P}(-1 \le X_1 \le 2) = \mathbf{P}(X_1 \in [-1,2]) &= \int_{[-1,2]} 1_{[0,1]}(x_1)\, d\mu_{L^1}(x_1) \\ &= \int_{\mathbb{R}} 1_{[-1,2]}(x_1)\, 1_{[0,1]}(x_1)\, d\mu_{L^1}(x_1) = \int_{\mathbb{R}} 1_{[-1,2]\cap[0,1]}(x_1)\, d\mu_{L^1}(x_1) \\ &= \int_{\mathbb{R}} 1_{[0,1]}(x_1)\, d\mu_{L^1}(x_1) = \mu_{L^1}([0,1]) = 1\end{aligned}$$

and

$$\begin{aligned}\mathbf{P}(-1 \leq X_2 \leq 1) = \mathbf{P}(X_2 \in [-1,1]) &= \int_{[-1,1]} \frac{1}{2} 1_{[0,2]}(x_2)\, d\mu_{L^1}(x_2) \\ &= \frac{1}{2}\int_{\mathbb{R}} 1_{[-1,1]}(x_2)\, 1_{[0,2]}(x_2)\, d\mu_{L^1}(x_2) = \int_{\mathbb{R}} 1_{[-1,1]\cap[0,2]}(x_2)\, d\mu_{L^1}(x_2) \\ &= \frac{1}{2}\int_{\mathbb{R}} 1_{[0,1]}(x_2)\, d\mu_{L^1}(x_2) = \frac{1}{2}\mu_{L^1}([0,1]) = \frac{1}{2}.\end{aligned}$$

It Then, follows

$$\mathbf{P}(-1 \leq X_1 \leq 2, -1 \leq X_2 \leq 1) = \mathbf{P}(-1 \leq X_1 \leq 2)\, \mathbf{P}(-1 \leq X_2 \leq 1).$$

□

To discuss the next problem we will exploit the following results

Theorem 724 (Change of Variable in Multiple Integrals) *Let O and Q be open subset of $\mathbb{R}^N$ and let $\Phi : O \to Q$ be a C^1-diffeomorphism. For any Riemann integrable function $h : D \to \mathbb{R}$, where $D \subseteq O$, we have*

$$\int_D h(x)dx = \int_{\Phi(D)} h(\Phi^{-1}(y)) \left|\det\left(J_{\Phi^{-1}}(y)\right)\right| dy$$

where $J_{\Phi^{-1}} : Q \to \mathbb{R}^{N\times N}$ is the Jacobean matrix of the inverse $\Phi^{-1} : Q \to O$ of Φ.

Corollary 725 *Let O and Q be open subset of $\mathbb{R}^N$ and let $\Phi : O \to Q$ be a C^1-diffeomorphism. Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space, and let $X : \Omega \to \mathbb{R}^N$ be an absolutely continuous random vector with density $f_X : \mathbb{R}^N \to \mathbb{R}_+$. Assume $X(\Omega) \subseteq O$. Then, for any $B \in \mathcal{B}(\mathbb{R}^N)$ such that $B \subseteq Q$ we have*

$$\mathbf{P}(\Phi(X) \in B) = \mathbf{P}\left(X \in \Phi^{-1}(B)\right) = \int_{\Phi^{-1}(B)} f_X(x)dx = \int_B f_X(\Phi^{-1}(y)) \left|\det\left(J_{\Phi^{-1}}(y)\right)\right| dy.$$

As a consequence, the random vector $\Phi \circ X : \Omega \to \mathbb{R}^N$ is absolutely continuous with density $f_{\Phi\circ X} : \mathbb{R}^N \to \mathbb{R}_+$ given by

$$f_{\Phi\circ X}(y) = f_X(\Phi^{-1}(y)) \left|\det\left(J_{\Phi^{-1}}(y)\right)\right| 1_{\Phi(X(\Omega))}(y), \quad \forall y \in \mathbb{R}^N.$$

Problem 726 *Let R and Θ be independent real random variables on a probability space Ω such that R is exponentially distributed with rate parameter λ and Θ is uniformly distributed in the interval $[0, 2\pi]$. Consider the random vector (U, V), where $U = R\cos(\Theta)$ and $V = R\sin(\Theta)$. Then, (U, V) is absolutely continuous and has joint density $f_{U,V} : \mathbb{R}^2 \to \mathbb{R}_+$ given by*

$$f_{U,V}(u, v) = \frac{\lambda e^{-\lambda\sqrt{u^2+v^2}}}{\sqrt{u^2+v^2}} 1_{\mathbb{R}\times(\mathbb{R}-\{0\})}(u, v),$$

for every $(u, v) \in \mathbb{R}^2$.

Solution. We know that the density function $f_R : \mathbb{R} \to \mathbb{R}$ [resp. $f_\Theta : \mathbb{R} \to \mathbb{R}$] of the random variable R [resp. Θ] can be written as

$$f_R(r) = \lambda e^{-\lambda r} 1_{(0,+\infty)}(r) \quad [\text{resp. } f_\Theta(\theta) = 1_{(0,2\pi)}(\theta)],$$

for every $r \in \mathbb{R}$ [resp. $\theta \in \mathbb{R}$]. Therefore, since R and Θ are independent, the joint density function $f_{R,\Theta} : \mathbb{R}^2 \to \mathbb{R}_+$ of R and Θ can be written as

$$f_{R,\Theta}(r,\theta) = f_R(r) f_\Theta(\theta) = \lambda e^{-\lambda r} 1_{(0,+\infty)}(r) 1_{(0,2\pi)}(\theta),$$

for every $(r,\theta) \in \mathbb{R}^2$. Now, consider the map $\Phi : (0,+\infty) \times [0,2\pi) \to \mathbb{R}^2 - \{(0,0)\}$ given by

$$\Phi(r,\theta) = (r\cos(\theta), r\sin(\theta)),$$

for every $(r,\theta) \in (0,+\infty)\times[0,2\pi)$. We have that Φ is invertible with inverse $\Phi^{-1} : \mathbb{R}^2 - \{(0,0)\} \to (0,+\infty) \times [0,2\pi)$ given by

$$\Phi^{-1}(u,v) = \left(\sqrt{u^2+v^2}, \theta(u,v)\right),$$

for every $(u,v) \in \mathbb{R}^2 - \{(0,0)\}$, where

$$\theta(u,v) \overset{\text{def}}{=} \begin{cases} \arctan\left(\frac{v}{u}\right) & \text{if } u > 0, v \geq 0 \\ \frac{\pi}{2} & \text{if } u = 0, v > 0 \\ \arctan\left(\frac{v}{u}\right) + \pi & \text{if } u < 0 \\ \frac{3\pi}{2} & \text{if } u = 0, v < 0 \\ \arctan\left(\frac{v}{u}\right) + 2\pi & \text{if } u > 0, v < 0 \end{cases}.$$

In addition, Φ is continuously differentiable on $(0,+\infty) \times (0,2\pi)$ with Jacobian matrix

$$J_\Phi(r,\theta) = \begin{pmatrix} \left(\frac{\partial \Phi_1}{\partial r}\right)_{(r,\theta)} & \left(\frac{\partial \Phi_1}{\partial \theta}\right)_{(r,\theta)} \\ \left(\frac{\partial \Phi_2}{\partial r}\right)_{(r,\theta)} & \left(\frac{\partial \Phi_2}{\partial \theta}\right)_{(r,\theta)} \end{pmatrix} = \begin{pmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{pmatrix},$$

such that

$$\det(J_\Phi(r,\theta)) = r > 0.$$

On the other hand, since we have

$$\left(\frac{\partial \theta}{\partial u}\right)_{(u,v)} = -\frac{v}{u^2+v^2} \quad \text{and} \quad \left(\frac{\partial \theta}{\partial v}\right)_{(u,v)} = \frac{u}{u^2+v^2},$$

for every $(u,v) \in \mathbb{R}^2 - \{(0,0)\}$ such that $u \neq 0$, we have that Φ^{-1} is continuously differentiable on $\mathbb{R}^2 - (\{0\} \times \mathbb{R})$ with Jacobian matrix

$$J_{\Phi^{-1}}(u,v) = \begin{pmatrix} \left(\frac{\partial \Phi_1^{-1}}{\partial u}\right)_{(u,v)} & \left(\frac{\partial \Phi_1^{-1}}{\partial v}\right)_{(u,v)} \\ \left(\frac{\partial \Phi_2^{-1}}{\partial u}\right)_{(u,v)} & \left(\frac{\partial \Phi_2^{-1}}{\partial v}\right)_{(u,v)} \end{pmatrix} = \begin{pmatrix} \frac{u}{\sqrt{u^2+v^2}} & \frac{v}{\sqrt{u^2+v^2}} \\ -\frac{v}{u^2+v^2} & \frac{u}{u^2+v^2} \end{pmatrix}.$$

such that

$$\det(J_{\Phi^{-1}}(u,v)) = \frac{1}{\sqrt{u^2+v^2}} > 0.$$

It follows that Φ is a C^1-diffeomorphism from $(0,+\infty)\times(0,2\pi)$ onto $\mathbb{R}^2-(\{0\}\times\mathbb{R})$. Considering that

$$\mu_L^2(\{0\}\times\mathbb{R})=0,$$

where $\mu_L^2:\mathcal{B}(\mathbb{R}^2)\to\mathbb{R}_+$ is the Borel-Lebesque measure on $\mathbb{R}^2$, by virtue of Theorem 724, the random vector

$$(U,V)\overset{\text{def}}{=}(R\cos(\Theta),R\sin(\Theta))=\Phi(R,\Theta)$$

is absolutely continuous with density $f_{U,V}:\mathbb{R}^2\to\mathbb{R}_+$ given by

$$\begin{aligned}f_{U,V}(u,v)&=f_{R,\Theta}\left(\Phi^{-1}(u,v)\right)\left|\det\left(J_{\Phi^{-1}}(u,v)\right)\right|1_{\mathbb{R}^2-(\{0\}\times\mathbb{R})}(u,v)\\&=\lambda e^{-\lambda\sqrt{u^2+v^2}}1_{(0,+\infty)}\left(\sqrt{u^2+v^2}\right)1_{(0,2\pi)}\left(\arctan\left(\frac{u}{v}\right)\right)\frac{1}{\sqrt{u^2+v^2}}1_{(\mathbb{R}-\{0\})\times\mathbb{R}}(u,v)\\&=\frac{\lambda e^{-\lambda\sqrt{u^2+v^2}}}{\sqrt{u^2+v^2}}1_{\mathbb{R}-\{0\}}(u),\end{aligned}$$

for every $(u,v)\in\mathbb{R}^2$.

Problem 727 *Let $(\Omega,\mathcal{E},\mathbf{P})\equiv\Omega$ be a probability space and let X and Y be real random variables on Ω with join density $f_{X,Y}:\mathbb{R}^2\to\mathbb{R}_+$ given by*

$$f_{X,Y}(x,y)\overset{def}{=}c\sqrt{x^2+y^2}1_{D(0;1)}(x,y),\quad\forall(x,y)\in\mathbb{R}^2,$$

where $c\in\mathbb{R}_+$ and $D(0;1)\equiv\{(x,y)\in\mathbb{R}^2:x^2+y^2\leq1\}$, that is the closed disk of $\mathbb{R}^2$ centered at $(0,0)\equiv0$ with radius 1.

Exercise 728 *1. Determine $c\in\mathbb{R}_+$ such that $f_{X,Y}$ is an actual density.*

2. Consider the real random variables $U\equiv|X|$ and $V\equiv Y/X$ and determine the joint density $f_{U,V}:\mathbb{R}^2\to\mathbb{R}_+$.

Solution.

1. By virtue of the properties of the Lebesgue integral, we have

$$\begin{aligned}\int_{\mathbb{R}^2}f_{X,Y}(x,y)\,d\mu_L^2(x,y)&=c\int_{\mathbb{R}^2}\sqrt{x^2+y^2}1_{D(0;1)}(x,y)\,d\mu_L^2(x,y)\\&=c\int_{D(0;1)}\sqrt{x^2+y^2}d\mu_L^2(x,y)\\&=c\int_{\tilde{B}(0;1)}\sqrt{x^2+y^2}d\mu_L^2(x,y)\\&=c\int_{\tilde{B}(0;1)}\sqrt{x^2+y^2}dxdy,\end{aligned}$$

where $\tilde{B}(0;1)\equiv B(0;1)-[0,1]\times\{0\}$ is the open disk of $\mathbb{R}^2$ centered at $(0,0)\equiv0$ with radius 1 deprived by the segment $[0,1]\times\{0\}$ such that

$$\mu_L^2\left(\tilde{B}(0;1)\right)=\mu_L^2(D(0;1)).$$

Hence, thanks to the change of coordinates $\Psi : \mathbb{R}_+ \times (0, 2\pi) \to \mathbb{R}^2$ given by

$$\Psi(r, \theta) \stackrel{\text{def}}{=} (r\cos(\theta), r\sin(\theta)), \quad \forall (r, \theta) \in \mathbb{R}_+ \times (0, 2\pi),$$

with Jacobian matrix

$$J_\Psi(r, \theta) \equiv \begin{pmatrix} \partial_r r\cos(\theta) & \partial_\theta r\cos(\theta) \\ \partial_r r\sin(\theta) & \partial_\theta r\sin(\theta) \end{pmatrix} = \begin{pmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{pmatrix},$$

such that

$$|\det(J_\Psi(r, \theta))| = r \qquad \text{and} \qquad \tilde{B}(0; 1) = \Psi((0, 1) \times (0, 2\pi)),$$

we obtain (see Theorem 724 where $\Phi^{-1} = \Psi$)

$$\begin{aligned} \int_{\tilde{B}(0;1)} \sqrt{x^2 + y^2} dxdy &= \int_{\Psi((0,1)\times(0,2\pi))} \sqrt{x^2 + y^2} dxdy \\ &= \int_{(0,1)\times(0,2\pi)} \sqrt{r^2\cos^2(\theta) + r^2\sin^2(\theta)} \, |\det(J_\Psi(r, \theta))| \, drd\theta \\ &= \int_{(0,1)\times(0,2\pi)} r^2 drd\theta \\ &= \int_{\theta=0}^{2\pi} \left(\int_{r=0}^{1} r^2 dr \right) d\theta = \frac{2\pi}{3}. \end{aligned}$$

It Then, follows that, to make $f_{X,Y}$ a density, we have to choose

$$c = \frac{3}{2\pi}.$$

2. First, let us observe that

$$\mathbf{P}(X = 0) = \mathbf{P}(X = 0, Y \in \mathbb{R}) = \int_{\{0\}\times\mathbb{R}} f_{X,Y}(x, y) \, d\mu_L(x, y) = 0.$$

Therefore, the random variables $U \equiv |X|$ and $V \equiv Y/X$ are both well defined. Furthermore

$$\mathbf{P}(U < 0) = 0.$$

Second, for every $x > 0$, we have

$$\begin{aligned} \mathbf{P}(U \leq x, V \leq y) &= \mathbf{P}(|X| \leq x, Y/X \leq y) \\ &= \mathbf{P}(X \leq x, Y/X \leq y) + \mathbf{P}(-X \leq x, Y/X \leq y). \end{aligned}$$

On the other hand

$$\mathbf{P}(U \leq x, V \leq y) = \int_{(0,x]\times(-\infty,y]} f_{U,V}(u, v) \, d\mu_L(u, v).$$

Hence, let us consider the changes of variables $\Phi_1 : \mathbb{R}_{++} \times \mathbb{R} \to \mathbb{R}_{++} \times \mathbb{R}$ and $\Phi_2 : \mathbb{R}_{--} \times \mathbb{R} \to \mathbb{R}_{++} \times \mathbb{R}$ given by

$$\Phi_1(x, y) \stackrel{\text{def}}{=} (x, y/x), \quad \forall (x, y) \in \mathbb{R}_{++}\times\mathbb{R} \quad \text{and} \quad \Phi_2(x, y) \stackrel{\text{def}}{=} (-x, y/x), \quad \forall (x, y) \in \mathbb{R}_{--}\times\mathbb{R},$$

having inverses $\Phi_1^{-1}:\mathbb{R}_{++}\times\mathbb{R}\to\mathbb{R}_{++}\times\mathbb{R}$ and $\Phi_2^{-1}:\mathbb{R}_{++}\times\mathbb{R}\to\mathbb{R}_{--}\times\mathbb{R}$ given by

$$\Phi_1^{-1}(u,v)\overset{\text{def}}{=}(u,uv),\quad\forall(u,v)\in\mathbb{R}_{++}\times\mathbb{R}\quad\text{and}\quad\Phi_2^{-1}(u,v)\overset{\text{def}}{=}(-u,-uv),\quad\forall(u,v)\in\mathbb{R}_{++}\times\mathbb{R},$$

with Jacobian matrices

$$J_{\Phi_1^{-1}}(u,v)\equiv\begin{pmatrix}\partial_u u & \partial_v u\\ \partial_u uv & \partial_v uv\end{pmatrix}=\begin{pmatrix}1 & 0\\ v & u\end{pmatrix}$$

and

$$J_{\Phi_2^{-1}}(u,v)\equiv\begin{pmatrix}\partial_u(-u) & \partial_v(-u)\\ \partial_u(-uv) & \partial_v(-uv)\end{pmatrix}=\begin{pmatrix}-1 & 0\\ -v & -u\end{pmatrix}.$$

We have

$$\det\left(J_{\Phi_1^{-1}}(u,v)\right)=u,\qquad\det\left(J_{\Phi_2^{-1}}(u,v)\right)=-u,$$

$$\mathbf{P}(X\leq x,Y/X\leq y)=\mathbf{P}(\Phi_1(X,Y)\in(0,x]\times(-\infty,y])=\mathbf{P}\left((X,Y)\in\Phi_1^{-1}((0,x]\times(-\infty,y])\right),$$

and

$$\mathbf{P}(-X\leq x,Y/X\leq y)=\mathbf{P}(\Phi_2(X,Y)\in(0,x]\times(-\infty,y])=\mathbf{P}\left((X,Y)\in\Phi_2^{-1}((0,x]\times(-\infty,y])\right).$$

Therefore,

$$\begin{aligned}\mathbf{P}(X,Y)\in\Phi_1^{-1}((0,x]\times(-\infty,y]) &= \int_{\Phi_1^{-1}((0,x]\times(-\infty,y])} f_{X,Y}(x,y)\,d\mu_L(x,y)\\ &= \int_{\Phi_1^{-1}((0,x]\times(-\infty,y])} f_{X,Y}(x,y)\,dudv\\ &= \int_{(0,x]\times(-\infty,y]} \left(f_{X,Y}\circ\Phi_1^{-1}\right)(u,v)\left|\det\left(J_{\Phi_1^{-1}}(u,v)\right)\right|dudv\\ &= \frac{3}{2\pi}\int_{(0,x]\times(-\infty,y]} \sqrt{u^2+u^2v^2}\,1_{D(0;1)}(u,uv)1_{\mathbb{R}_{++}}(u)\,ududv\\ &= \frac{3}{2\pi}\int_{(0,x]\times(-\infty,y]} u^2\sqrt{1+v^2}\,1_{D(0;1)}(u,uv)1_{\mathbb{R}_{++}}(u)\,dudv.\end{aligned}$$

and

$$\begin{aligned}\mathbf{P}(X,Y)\in\Psi_2((0,x)\times(-\infty,y)) &= \int_{\Phi_2^{-1}((0,x]\times(-\infty,y])} f_{X,Y}(x,y)\,d\mu_L(x,y)\\ &= \int_{\Phi_2^{-1}((0,x]\times(-\infty,y])} f_{X,Y}(x,y)\,dxdy\\ &= \int_{(0,x]\times(-\infty,y]} \left(f_{X,Y}\circ\Phi_2^{-1}\right)(u,v)\left|\det\left(J_{\Phi_2^{-1}}(u,v)\right)\right|dudv\\ &= \frac{3}{2\pi}\int_{(0,x]\times(-\infty,y]} \sqrt{u^2+u^2v^2}\,1_{D(0;1)}(-u,-uv)1_{\mathbb{R}_{++}}(u)\,ududv\\ &= \frac{3}{2\pi}\int_{(0,x]\times(-\infty,y]} u^2\sqrt{1+v^2}\,1_{D(0;1)}(-u,-uv)1_{\mathbb{R}_{++}}(u)\,dudv.\end{aligned}$$

Now,

$$1_{D(0;1)}(u,uv)1_{\mathbb{R}_{++}}(u) = 1_{D(0;1)}(-u,-uv)1_{\mathbb{R}_{++}}(u) = \begin{cases} 1 & \text{if } u > 0^2 \text{ and } (1+v^2) \leq 1 \\ 0 & \text{otherwise} \end{cases}.$$

We Then, have

$$\int_{(0,x]\times(-\infty,y]} f_{U,V}(u,v)\, d\mu_L(u,v) = \frac{3}{\pi}\int_{(0,x]\times(-\infty,y]} u^2\sqrt{1+v^2}1_{\{u>0,u^2(1+v^2)\leq 1\}}(u,v)dudv,$$

which yields

$$f_{U,V}(u,v) = \frac{3}{\pi}u^2\sqrt{1+v^2}1_{\{u>0,u^2(1+v^2)\leq 1\}}(u,v),$$

for every $(u,v) \in \mathbb{R}^2$. □

6.3 Moments of a Random Vector

Let $X : \Omega \to \mathbb{R}^N$ be an $\mathcal{E}$-random vector with entries $X_1, \ldots, X_N$, and let $\|\cdot\|_2 : \mathbb{R}^N \to \mathbb{R}_+$ be the Euclidean norm on $\mathbb{R}^N$ (see Definition 14).

Remark 729 *The map $\|\cdot\|_2 \circ X : \Omega \to \mathbb{R}_+$, briefly denoted by $\|X\|_2$, given by*

$$\|\cdot\|_2 \circ X(\omega) \overset{def}{=} \|X(\omega)\|_2, \quad \forall \omega \in \Omega,$$

is a positive $\mathcal{E}$-random variable.

Definition 730 *Fixed any $p > 0$, we say that X has* finite moment *of order p or* finite expectation, *if*

$$\int_\Omega \|X\|_2^p \, d\mathbf{P} < \infty. \tag{6.37}$$

Proposition 731 *The random vector $X : \Omega \to \mathbb{R}^N$ has finite moment of order 1 if and only if all entries $X_1, \ldots, X_N$ have. In symbols*

$$\int_\Omega \|X\|_2 \, d\mathbf{P} < \infty \Leftrightarrow \int_\Omega |X_K| \, d\mathbf{P} < \infty, \quad \forall K = 1, \ldots, N. \tag{6.38}$$

Proof. Following the proof of Proposition 22, we can write

$$|X_K| = \left(X_K^2\right)^{1/2} \leq \left(\sum_{K=1}^N X_K^2\right)^{1/2} = \|X\|_2,$$

for every $K = 1, \ldots, N$. Therefore, from the integrability of $\|X\|$ on Ω, it follows the integrability of all entries $X_1, \ldots, X_N$. Conversely, if all entries $X_1, \ldots, X_N$ are integrable on Ω, from

$$\left(\sum_{K=1}^N X_K^2\right) \leq \left(\sum_{K=1}^N |X_K|\right)^2,$$

it follows

$$\|X\|_2 = \left(\sum_{K=1}^N X_K^2\right)^{1/2} \leq \sum_{K=1}^N |X_K|,$$

which implies that $\|X\|$ is integrable on Ω. □

Proposition 732 *The random vector $X : \Omega \to \mathbb{R}^N$ has finite moment of order 2 if and only if all products $X_J X_K$ of the entries of X, on vaying of $J, K = 1, \ldots, N$, have finite moments of order 1. In symbols*

$$\int_\Omega \|X\|_2^2 \, d\mathbf{P} < \infty \Leftrightarrow \int_\Omega |X_J X_K| \, d\mathbf{P} < \infty, \quad \forall J, K = 1, \ldots, N. \tag{6.39}$$

Proof. Assume $\|X\|_2^2$ is integrable on Ω. Since

$$X_K^2 \leq \left(\sum_{K=1}^N X_K^2 \right) = \|X\|_2^2 ,$$

for every $K = 1, \ldots, N$, we obtain that all entries $X_1, \ldots, X_N$ of X have finite moment of order 2. From the Cauchy Schwarz inequality (see Proposition 661) it Then, follows that all products $X_J X_K$ of the entries of X, on vaying of $J, K = 1, \ldots, N$, have finite moments of order 1. Conversely, if all products $X_J X_K$ of the entries of X, on varying of $J, K = 1, \ldots, N$, have finite moments of order 1, in particular all squares $X_1^2, \ldots, X_N^2$ of the entries of X have finite moment of order 1. This clearly implies that also $\|X\|_2^2$ has finite moment of order 1. □

Proposition 733 *Given any $K \in \mathbb{N}$, write*

$$\mathbb{N}_0^N (K) \equiv \left\{ (K_1, \ldots, K_N) \in \mathbb{N}_0^N : \sum_{j=1}^N K_j = K \right\}.$$

The N variate real random variable X has finite Kth moment if and only if all products $X_1^{K_1} X_2^{K_2} \cdots X_N^{K_N}$ of the entries of X, on varying of $(K_1, \ldots, K_N) \in \mathbb{N}_0^N (K)$, have finite moments of order 1. In symbols

$$\int_\Omega \|X\|_2^K \, d\mathbf{P} < \infty \Leftrightarrow \int_\Omega \left| X_1^{K_1} X_2^{K_2} \cdots X_N^{K_N} \right| d\mathbf{P} < \infty, \quad \forall \, (K_1, \ldots, K_N) \in \mathbb{N}_0^N (K) . \tag{6.40}$$

Proof. . □

Proposition 734 *If X has* finite moment *of order p, for some $p > 0$, Then, X has finite moment of order p' for every $0 < p' \leq p$.*

Proof. For any $p' > 0$, the events $\left\{ \|X\|_2^{p'} \leq 1 \right\}$ and $\left\{ \|X\|_2^{p'} > 1 \right\}$ constitute a partition of Ω . Considering the positivity of the $\mathcal{E}$-random variable $\|X\|_2^{p'}$, we can then write

$$\int_\Omega \|X\|_2^{p'} \, d\mathbf{P} = \int_{\left\{ \|X\|_2^{p'} \leq 1 \right\}} \|X\|_2^{p'} \, d\mathbf{P} + \int_{\left\{ \|X\|_2^{p'} > 1 \right\}} \|X\|_2^{p'} \, d\mathbf{P}$$

Now, we have

$$\int_{\left\{ \|X\|_2^{p'} \leq 1 \right\}} \|X\|_2^{p'} \, d\mathbf{P} \leq \mathbf{P} \left(\|X\|_2^{p'} \leq 1 \right) \leq 1.$$

On the other hand, since on the occurrence of the event $\left\{ \|X\|_2^{p'} > 1 \right\}$ and for $p' < p$ we have

$$\|X\|_2^{p'} \leq \|X\|_2^p ,$$

we also have

$$\int_{\left\{\|X\|_2^{p'}>1\right\}} \|X\|_2^{p'}\, d\mathbf{P} \leq \int_{\left\{\|X\|_2^{p'}>1\right\}} \|X\|_2^{p}\, d\mathbf{P} \leq \int_{\Omega} \|X\|_2^{p}\, d\mathbf{P}.$$

Therefore, if Equation (6.37) holds true, we have that the same equation holds true for every $p' < p$. □

Definition 735 *In case $X : \Omega \to \mathbb{R}^N$ has finite moment of order 1, we call* expectation *or* mean *or also* moment of order 1 *of X the N-dimensional real vector*

$$\mathbf{E}[X] \equiv \mu_X \overset{def}{=} (\mathbf{E}[X_1], \ldots, \mathbf{E}[X_N])^{\intercal} \equiv (\mu_{X_1}, \ldots, \mu_{X_N}).$$

given by

$$\mathbf{E}[X_K] \equiv \mu_{X_K} \overset{def}{=} \int_{\Omega} X_K d\mathbf{P}, \quad \forall K = 1, \ldots, N.$$

Notation 736 *We denote by $\mathcal{L}^1(\Omega; \mathbb{R}^N)$ the set of all random vectors on Ω with states in $\mathbb{R}^N$ having finite moment of order 1.*

Theorem 737 *Theorem the set $\mathcal{L}^1(\Omega; \mathbb{R}^N)$ is a linear space and the maps*

$$|\cdot|_1 : \mathcal{L}^1(\Omega; \mathbb{R}^N) \to \mathbb{R}_+, \quad |X|_1 \overset{def}{=} \left(\sum_{K=1}^{N} \mathbf{E}\left[|X_K|\right]^2 \right)^{1/2}, \quad \forall X \in \mathcal{L}^1(\Omega; \mathbb{R}^N),$$

$$|\cdot|_{1\ sum} : \mathcal{L}^1(\Omega; \mathbb{R}^N) \to \mathbb{R}_+, \quad |X|_{1\ sum} \overset{def}{=} \sum_{K=1}^{N} \mathbf{E}\left[|X_K|\right], \quad \forall X \in \mathcal{L}^1(\Omega; \mathbb{R}^N),$$

$$|\cdot|_{1\ max} : \mathcal{L}^1(\Omega; \mathbb{R}^N) \to \mathbb{R}_+, \quad |X|_{1\ max} \overset{def}{=} \max_{K=1,\ldots,N} \left\{\mathbf{E}\left[|X_K|\right]\right\}, \quad \forall X \in \mathcal{L}^1(\Omega; \mathbb{R}^N),$$

are equivalent semi-norms on $\mathcal{L}^1(\Omega; \mathbb{R}^N)$.

Definition 738 *In case $X : \Omega \to \mathbb{R}^N$ has finite moment of order 2, we call* variance-covariance *of X the N-order real matrix*

$$Var(X) \equiv \Sigma_X^2 \equiv (\sigma_{J,K})_{J,K=1}^{N}$$

given by

$$\sigma_{J,K} \overset{def}{=} Cov(X_J, X_K) \equiv \mathbf{E}\left[(X_J - \mathbf{E}[X_J])(X_K - \mathbf{E}[X_K])\right], \quad \forall J, K = 1, \ldots, N.$$

Proposition 739 *In case $X : \Omega \to \mathbb{R}^N$ has finite moment of order 2, the variance-covariance matrix $Var(X)$ is symmetric and positive semidefinite.*

Proof. The symmetry being evident, we prove only the positive semidefiniteness. Thanks to the linearity of the expectation operator, we have

$$\begin{aligned} u^{\intercal} Var\left(X\right) u &\equiv \sum_{J,K=1}^{N} u_J Cov\left(X_J, X_K\right) u_K \\ &\equiv \sum_{J,K=1}^{N} u_J \mathbf{E}\left[\left(X_J - \mathbf{E}\left[X_J\right]\right)\left(X_K - \mathbf{E}\left[X_K\right]\right)\right] u_K \\ &= \mathbf{E}\left[\sum_{J,K=1}^{N} u_J \left(X_J - \mathbf{E}\left[X_J\right]\right) u_K \left(X_K - \mathbf{E}\left[X_K\right]\right)\right] \\ &= \mathbf{E}\left[\left(\sum_{J=1}^{N} u_J \left(X_J - \mathbf{E}\left[X_J\right]\right)\right)\left(\sum_{K=1}^{N} u_K \left(X_K - \mathbf{E}\left[X_K\right]\right)\right)\right] \\ &= \mathbf{E}\left[\left(\sum_{K=1}^{N} u_K \left(X_K - \mathbf{E}\left[X_K\right]\right)\right)^2\right] \geq 0, \end{aligned}$$

for every $u \in \mathbb{R}^N$, as desired. □

Notation 740 *We denote by* $\mathcal{L}^2(\Omega; \mathbb{R}^N)$ *the set of all random vectors on* Ω *with states in* $\mathbb{R}^N$ *having finite moment of order* 2*.*

Let $X, Y \in \mathcal{L}^2(\Omega; \mathbb{R}^N)$, $X \equiv (X_1, \ldots, X_N)$, $Y \equiv (Y_1, \ldots, Y_N)$.

Lemma 741 *The random variable*

$$X^{\intercal} Y \equiv \sum_{K=1}^{N} X_K Y_K$$

has finite moment of order 1*.*

Proof. By virtue of the Schwarz inequality, we have

$$\mathbf{E}\left[\left|X_J Y_K\right|\right] \leq \mathbf{E}\left[X_J^2\right]^{1/2} \mathbf{E}\left[Y_K^2\right]^{1/2},$$

for all $J, K = 1, \ldots, N$. On the other hand, the random vectors X and Y are assumed to have autocovariances. The desired result immediately follows. □

Theorem 742 *Theorem the set* $\mathcal{L}^2(\Omega; \mathbb{R}^N)$ *is a linear space, the maps*

$$\left|\cdot\right|_2 : \mathcal{L}^2(\Omega; \mathbb{R}^N) \to \mathbb{R}_+, \quad \left|X\right|_2 \overset{def}{=} \left(\sum_{K=1}^{N} \mathbf{E}\left[X_K^2\right]\right)^{1/2}, \quad \forall X \in \mathcal{L}^2(\Omega; \mathbb{R}^N),$$

$$\left|\cdot\right|_{2\ sum} : \mathcal{L}^2(\Omega; \mathbb{R}^N) \to \mathbb{R}_+, \quad \left|X\right|_{2\ sum} \overset{def}{=} \sum_{K=1}^{N} \mathbf{E}\left[X_K^2\right], \quad \forall X \in \mathcal{L}^2(\Omega; \mathbb{R}^N),$$

$$\left|\cdot\right|_{2\ max} : \mathcal{L}^2(\Omega; \mathbb{R}^N) \to \mathbb{R}_+, \quad \left|X\right|_{2\ max} \overset{def}{=} \max_{K=1,\ldots,N} \left\{\mathbf{E}\left[X_K^2\right]\right\}, \quad \forall X \in \mathcal{L}^2(\Omega; \mathbb{R}^N),$$

are equivalent semi-norms on $\mathcal{L}^2(\Omega;\mathbb{R}^N)$*, and the map*

$$(\cdot,\cdot):\mathcal{L}^2(\Omega;\mathbb{R}^N)\to\mathbb{R}_+,\quad (X,Y)\overset{def}{=}\mathbf{E}\left[X^\intercal Y\right],\quad \forall X,Y\in\mathcal{L}^2(\Omega;\mathbb{R}^N)$$

is a symmetric bilinear form on $\mathcal{L}^2(\Omega;\mathbb{R}^N)$ *such that*

$$(X,X)^{1/2}=|X|_2^2,\quad \forall X\in\mathcal{L}^2(\Omega;\mathbb{R}^N).$$

Proof. . □

Let $Y:\Omega\to\mathbb{R}^M$ [resp. $Y:\Omega\to\mathbb{R}^N$] be an $\mathcal{E}$-random vector on Ω with states in $\mathbb{R}^M$ [resp. $\mathbb{R}^N$], $X\equiv(X_1,\ldots,X_N)^\intercal$ [resp. $Y\equiv(Y_1,\ldots,Y_N)^\intercal$].

Lemma 743 *Assume* $X\in\mathcal{L}^2(\Omega;\mathbb{R}^M)$ *and* $Y\in\mathcal{L}^2(\Omega;\mathbb{R}^N)$*. Then, the random variables*

$$X_JY_K$$

have finite moment of order 1 *for every* $J=1,\ldots,M$ *and* $K=1,\ldots,N$*.*

Definition 744 *In case* $X\in\mathcal{L}^2(\Omega;\mathbb{R}^M)$ *and* $Y\in\mathcal{L}^2(\Omega;\mathbb{R}^N)$*, we call* covariance *of* X *and* Y *the* $M\times N$ *real matrix*

$$Cov\left(X,Y\right)\equiv\left(\sigma_{J,K}\right)_{J=1,\ldots,J,\ K=1,\ldots,N},$$

where

$$\sigma_{J,K}\overset{def}{=}Cov\left(X_J,Y_K\right),\quad \forall J=1,\ldots,N,\ K=1,\ldots,N.$$

Chapter 7

Independent Random Variables

The notion of σ-algebra generated by a random variable allows to transfer to random variables the definition of independence already introduced for events. Since random variables represent quantitative or categorical observations of a random phenomenon, the goal is to capture the idea that the possible outcomes of these observations are independent. More precisely, two groups of random variables will be thought independent if the probability of observing any set of states taken by one group does not influence the probability of observing any set of states of the other group.

7.1 Pairs of Independent Real Random Variables

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space, let $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ the real Borel state space, let X and Y be real $\mathcal{E}$-random variables on Ω, and let $\sigma(X)$ and $\sigma(Y)$ be the sub-σ-algebras of $\mathcal{E}$ generated by X and Y, respectively.

Definition 745 *We say that X and Y are independent if $\sigma(X)$ and $\sigma(Y)$ are independent (see Definition 377), that is*

$$\mathbf{P}(E \cap F) = \mathbf{P}(E)\,\mathbf{P}(F), \tag{7.1}$$

for every $E \in \sigma(X)$ and every $F \in \sigma(Y)$. Equivalently, if

$$\mathbf{P}(X \in B, Y \in C) = \mathbf{P}(X \in B)\,\mathbf{P}(Y \in C), \tag{7.2}$$

for all $B, C \in \mathcal{B}(\mathbb{R})$.

Example 746 *If $X_0 : \Omega \to \mathbb{R}$ is a Dirac random variable concentrated in some $x_0 \in \mathbb{R}$, then X_0 is independent of any random variable Y.*

Discussion. We have

$$\{X_0 \in B\} = \begin{cases} \Omega, & \text{if } x_0 \in B, \\ \varnothing, & \text{if } x_0 \notin B, \end{cases}$$

for every $B \in \mathcal{B}(\mathbb{R})$. Therefore, given any $C \in \mathcal{B}(\mathbb{R})$, we have

$$\mathbf{P}(X_0 \in B, Y \in C) = \mathbf{P}(\Omega, Y \in C) = \mathbf{P}(Y \in C) = \mathbf{P}(\Omega)\,\mathbf{P}(Y \in C) = \mathbf{P}(X_0 \in B)\,\mathbf{P}(Y \in B)$$

for every $B \in \mathcal{B}(\mathbb{R})$ such that $x_0 \in B$ and

$$\mathbf{P}(X_0 \in B, Y \in N) = \mathbf{P}(\varnothing, Y \in N) = \mathbf{P}(\varnothing) = \mathbf{P}(\varnothing)\,\mathbf{P}(Y \in N) = \mathbf{P}(X_0 \in B)\,\mathbf{P}(Y \in N)$$

for every $B \in \mathcal{B}(\mathbb{R})$ such that $x_0 \notin B$. Hence, Equation (7.2) holds true for all $B, C \in \mathcal{B}(\mathbb{R})$. This proves the claim. □

Example 747 *If $X_0 : \Omega \to \mathbb{R}$ is a Dirac random variable concentrated in some $x_0 \in \mathbb{R}$, then X_0 is independent of itself. Moreover, it is the only real random variable which is independent of itself.*

Discussion. The discussion of Example 746 clearly applies also when $Y = X_0$. Hence, X_0 is independent of itself. To show that X_0 is the only real random variable which is independent of itself, assume that X is not a Dirac real random variable. Then, there exist at least two distinct values taken by X, say x_1 and x_2, and two Borel sets, say B_1 and B_1, such that

$$x_1 \in B_1, \quad x_2 \in B_2, \quad B_2 = B_1^c, \quad \mathbf{P}(X \in B_1) > 0, \quad \mathbf{P}(X \in B_2) > 0.$$

Then, we have

$$\mathbf{P}(X \in B_1, X \in B_2) = \mathbf{P}(X \in B_1 \cap B_2) = \mathbf{P}(X \in \varnothing) = \mathbf{P}(\varnothing) = 0$$

and

$$\mathbf{P}(X = x_1)\,\mathbf{P}(X = x_2) > 0.$$

This prevents X from being independent of itself. □

Example 748 *Let R be Rademacher distributed, $R \sim Rad(p)$, for some $p \in (0,1)$. Then, the random variables R and R^2 are independent.*

Discussion. It is sufficient to observe that the random variable R^2 is a Dirac random variable concentrated in 1. □

Proposition 749 *Let X and Y the indicator functions of some events E and F, respectively. In symbols $X = 1_E$ and $Y = 1_F$, where $E, F \in \mathcal{E}$. Then, the random variables X and Y are independent if and only if the events E and F are.*

Proof. Consider any $B, C \in \mathcal{B}(\mathbb{R})$ we have the following possible sixteen cases

$$\begin{array}{llllllll}
01) & 0 \in B & 1 \in B & 0 \in C & 1 \in C & \Rightarrow & \{1_E \in B\} = \Omega & \{1_F \in C\} = \Omega \\
02) & 0 \in B & 1 \in B & 0 \in C & 1 \notin C & \Rightarrow & \{1_E \in B\} = \Omega & \{1_F \in C\} = F^c \\
03) & 0 \in B & 1 \in B & 0 \notin C & 1 \in C & \Rightarrow & \{1_E \in B\} = \Omega & \{1_F \in C\} = F \\
04) & 0 \in B & 1 \in B & 0 \notin C & 1 \notin C & \Rightarrow & \{1_E \in B\} = \Omega & \{1_F \in C\} = \varnothing \\
05) & 0 \in B & 1 \notin B & 0 \in C & 1 \in C & \Rightarrow & \{1_E \in B\} = E^c & \{1_F \in C\} = \Omega \\
06) & 0 \in B & 1 \notin B & 0 \in C & 1 \notin C & \Rightarrow & \{1_E \in B\} = E^c & \{1_F \in C\} = F^c \\
07) & 0 \in B & 1 \notin B & 0 \notin C & 1 \in C & \Rightarrow & \{1_E \in B\} = E^c & \{1_F \in C\} = F \\
08) & 0 \in B & 1 \notin B & 0 \notin C & 1 \notin C & \Rightarrow & \{1_E \in B\} = E^c & \{1_F \in C\} = \varnothing \\
09) & 0 \notin B & 1 \in B & 0 \in C & 1 \in C & \Rightarrow & \{1_E \in B\} = E & \{1_F \in C\} = \Omega \\
10) & 0 \notin B & 1 \in B & 0 \in C & 1 \notin C & \Rightarrow & \{1_E \in B\} = E & \{1_F \in C\} = F^c \\
11) & 0 \notin B & 1 \in B & 0 \notin C & 1 \in C & \Rightarrow & \{1_E \in B\} = E & \{1_F \in C\} = F \\
12) & 0 \notin B & 1 \in B & 0 \notin C & 1 \notin C & \Rightarrow & \{1_E \in B\} = E & \{1_F \in C\} = \varnothing \\
13) & 0 \notin B & 1 \notin B & 0 \in C & 1 \in C & \Rightarrow & \{1_E \in B\} = \varnothing & \{1_F \in C\} = \Omega \\
14) & 0 \notin B & 1 \notin B & 0 \in C & 1 \notin C & \Rightarrow & \{1_E \in B\} = \varnothing & \{1_F \in C\} = F^c \\
15) & 0 \notin B & 1 \notin B & 0 \notin C & 1 \in C & \Rightarrow & \{1_E \in B\} = \varnothing & \{1_F \in C\} = F \\
16) & 0 \notin B & 1 \notin B & 0 \notin C & 1 \notin C & \Rightarrow & \{1_E \in B\} = \varnothing & \{1_F \in C\} = \varnothing
\end{array}.$$

According to the above table, Equation (7.2) is clearly fulfilled in cases 01), 02), 03), 04), 05), 08), 09), 12), 13), 14), 15), 16). In addition, cases 07) and 10) are symmetric. Therefore, we are left to consider the cases 06), 07), 11) in which Equation (7.2) becomes

$$\mathbf{P}\left(E^c \cap F^c\right) = \mathbf{P}\left(E^c\right)\mathbf{P}\left(F^c\right), \quad \mathbf{P}\left(E^c \cap F\right) = \mathbf{P}\left(E^c\right)\mathbf{P}\left(F\right), \quad \mathbf{P}\left(E \cap F\right) = \mathbf{P}\left(E\right)\mathbf{P}\left(F\right), \tag{7.3}$$

correspondingly. Now, under the assumption of independence of E, F, on account of Proposition 366, Equation (7.3) holds true. Hence, we can conclude that (7.2) is fulfilled for all $B, C \in \mathcal{B}(\mathbb{R})$. This yields the indepencence of the random variables X and Y. Conversely, assume that X and Y are independent. As a particular case of (7.2), we have

$$\mathbf{P}\left(X = 1, Y = 1\right) = \mathbf{P}\left(X = 1\right)\mathbf{P}\left(Y = 1\right). \tag{7.4}$$

On the other hand,

$$\{X = 1\} \equiv \{1_E = 1\} = E \quad \text{and} \quad \{Y = 1\} \equiv \{1_F = 1\} = F.$$

Hence, (7.4) yields the independence of E and F. □

Let $\mathcal{B}$ be a basis for the Borel σ-algebra $\mathcal{B}(\mathbb{R})$.

Proposition 750 *The random variables X and Y are independent if and only if we have*

$$\mathbf{P}\left(X \in B, Y \in C\right) = \mathbf{P}\left(X \in B\right)\mathbf{P}\left(Y \in C\right), \tag{7.5}$$

for all $B, C \in \mathcal{B}$.

Proof. . □

Let $\mathcal{B}$ be a basis for the σ-algebra $\mathcal{B}(\mathbb{R})$, let $P_X : \mathcal{B}(\mathbb{R}) \to \mathbb{R}_+$ and $P_Y : \mathcal{B}(\mathbb{R}) \to \mathbb{R}_+$ be the distributions of the random variables X and Y, respectively, and let $P_{X,Y} : \mathcal{B}(\mathbb{R}^2) \to \mathbb{R}_+$ be the distribution of the random vector (X, Y).

Proposition 751 *The random variables X and Y are independent if and only if we have*

$$P_{X,Y}\left(B \times C\right) = P_X\left(B\right)P_Y\left(C\right), \tag{7.6}$$

for all $B, C \in \mathcal{B}$.

Proof. The claim immediately follows from Proposition 750 by observing that

$$P_{X,Y}\left(B \times C\right) = \mathbf{P}\left(X \in B, Y \in C\right)$$

and

$$P_X\left(B\right) = \mathbf{P}\left(X \in B\right), \quad P_Y\left(C\right) = \mathbf{P}\left(Y \in C\right).$$

for all $B, C \in \mathcal{B}$. □

Proposition 752 *The random variables X and Y are independent if and only if we have*

$$P_{X,Y} = P_X \otimes P_Y, \tag{7.7}$$

where $\otimes$ is the tensor product of measures.

Proof. . □

Proposition 753 *Assume X and Y are independent and $X, Y \in L^1(\Omega;\mathbb{R})$. Then, the product random variable XY has finite moment of order 1 and we have*

$$\mathbf{E}[XY] = \mathbf{E}[X]\,\mathbf{E}[Y]. \tag{7.8}$$

As a consequence, X and Y are uncorrelated, that is

$$Cov(X,Y) = 0. \tag{7.9}$$

Proof. Since $X, Y \in L^1(\Omega;\mathbb{R})$, we have

$$\int_{\mathbb{R}} |x|\, dP_X < \infty \quad \text{and} \quad \int_{\mathbb{R}} |y|\, dP_Y < \infty.$$

This implies

$$\int_{\mathbb{R}}\int_{\mathbb{R}} |xy|\, dP_X dP_Y = \int_{\mathbb{R}} |x|\, dP_X \int_{\mathbb{R}} |y|\, dP_Y < \infty.$$

On the other hand, since X and Y are assumed to be independent, by virtue of the Equation (**??**), we have

$$\int_{\mathbb{R}}\int_{\mathbb{R}} |xy|\, dP_{X,Y} = \int_{\mathbb{R}}\int_{\mathbb{R}} |xy|\, dP_X dP_Y.$$

It follows that $XY \in L^1(\Omega;\mathbb{R})$. Moreover, thanks to the Fubini theorem, we have

$$\begin{aligned}\mathbf{E}[XY] &= \int_\Omega XY d\mathbf{P} = \int_{\mathbb{R}^2} xy dP_{X,Y} = \int_{\mathbb{R}^2} xy dP_X dP_Y = \int_{\mathbb{R}} x dP_X \int_{\mathbb{R}} y dP_Y = \int_\Omega X d\mathbf{P} \int_\Omega Y d\mathbf{P} \\ &= \mathbf{E}[X]\,\mathbf{E}[Y],\end{aligned}$$

as claimed in (7.8). □

Example 754 *Let X be standard Gaussian distributed, $X \sim N(0,1)$. Then, the random variables X and X^2 are uncorrelated but not independent.*

Discussion. Since $X \sim N(0,1)$, we have

$$\mathbf{E}\left[X^3\right] = 0.$$

Hence,

$$Cov\left(X, X^2\right) = \mathbf{E}\left[X^3\right] - \mathbf{E}\left[X^2\right]\mathbf{E}[X] = 0.$$

This shows that X and X^2 are uncorrelated. On the other hand, since

$$\{X^2 \geq 1/4\} = \{X \leq -1/2\} \cup \{X \geq 1/2\} \quad \text{and} \quad \{X \leq -1/2\} \cap \{X \geq 1/2\} = \varnothing,$$

we have

$$\begin{aligned}\mathbf{P}\left(X \geq 1/2, X^2 \geq 1/4\right) &= \mathbf{P}\left(\{X \geq 1/2\} \cap (\{X \leq -1/2\} \cup \{X \geq 1/2\})\right) \\ &= \mathbf{P}\left(\{X \geq 1/2\} \cap \{X \leq -1/2\}\right) + \mathbf{P}\left(\{X \geq 1/2\} \cap \{X \geq 1/2\}\right) \\ &= \mathbf{P}(X \geq 1/2).\end{aligned}$$

Therefore, since

$$\mathbf{P}\left(X^2 \geq 1/4\right) < 1,$$

we obtain

$$\mathbf{P}\left(X \geq 1/2, X^2 \geq 1/4\right) < \mathbf{P}\left(X \geq 1/2\right)\mathbf{P}\left(X^2 \geq 1/4\right).$$

This prevents X and X^2 to be independent (see also Example 760). □

Example 755 *Let X be standard Gaussian distributed, $X \sim N(0,1)$. Consider the random variable Y_c given by*

$$Y_c \overset{def}{=} \begin{cases} X, & |X| \leq c, \\ -X, & |X| > c. \end{cases}$$

where $c > 0$. We have that $Y_c \sim N(0,1)$ for every $c > 0$. In addition, there exists $c_0 > 0$ such that the random variables X and Y_{c_0} are uncorrelated, but not independent.

Discussion. Using the standard symbol Φ to denote the distribution function of $X \sim N(0,1)$, consider the distribution function F_Y of Y. Thanks to the properties of $N(0,1)$, we have

$$\begin{aligned} F_{Y_c}(y) &= \mathbf{P}(Y_c \leq y) \\ &= \mathbf{P}(Y_c \leq y, |X| \leq c) + \mathbf{P}(Y_c \leq y, |X| > c) \\ &= \mathbf{P}(Y_c \leq y \mid |X| \leq c)\,\mathbf{P}(|X| \leq c) + \mathbf{P}(Y_c \leq y \mid |X| > c)\,\mathbf{P}(|X| > c) \\ &= \mathbf{P}(X \leq y \mid |X| \leq c)\,\mathbf{P}(|X| \leq c) + \mathbf{P}(-X \leq y \mid |X| > c)\,\mathbf{P}(|X| > c) \\ &= \mathbf{P}(X \leq y, |X| \leq c) + \mathbf{P}(X \geq -y, |X| > c) \\ &= \mathbf{P}(X \leq y, |X| \leq c) + \mathbf{P}(X \leq y, |X| > c) \\ &= \mathbf{P}(X \leq y) \\ &= \Phi(y), \end{aligned}$$

for every $y \in \mathbb{R}$. This proves that $Y_c \sim N(0,1)$ for every $c > 0$. Now, we have

$$|XY_c| = \begin{cases} |XX| = X^2, & |X| \leq c, \\ |X(-X)| = X^2, & |X| > c. \end{cases}$$

Hence, $XY_c \in L^1(\Omega;\mathbb{R})$ and

$$\begin{aligned} \mathbf{E}[XY_c] &= \int_\Omega XY_c d\mathbf{P} = \int_{|X|\leq c} XY_c d\mathbf{P} + \int_{|X|>c} XY_c d\mathbf{P} \\ &= \int_{|X|\leq c} X^2 d\mathbf{P} - \int_{|X|>c} X^2 d\mathbf{P}, \end{aligned}$$

where

$$\int_{|X|\leq c} X^2 d\mathbf{P} = \sqrt{\frac{2}{\pi}} \int_0^c x^2 e^{-\frac{x^2}{2}} dx \quad \text{and} \quad \int_{|X|>c} X^2 d\mathbf{P} = \sqrt{\frac{2}{\pi}} \int_c^\infty x^2 e^{-\frac{x^2}{2}} dx.$$

Both the integrals are clearly continuous on varying of $c > 0$. In addition,

$$\lim_{c\to 0^+} \mathbf{E}[XY_c] = -\int_\Omega X^2 d\mathbf{P} = -\mathbf{E}\left[X^2\right] = -1$$

and

$$\lim_{c\to +\infty} \mathbf{E}[XY_c] = \int_\Omega X^2 d\mathbf{P} = \mathbf{E}\left[X^2\right] = 1.$$

It follows, there exists $c_0 > 0$ such that

$$\mathbf{E}[XY_{c_0}] = 0.$$

For such a c_0 the standard normal random variables X and Y_{c_0} are clearly uncorrelated. In the end, we have

$$\begin{aligned}
&\mathbf{P}\left(X \leq x, Y_{c_0} \leq y\right) \\
&= \mathbf{P}\left(X \leq x, Y_{c_0} \leq y, |X| \leq c_0\right) + \mathbf{P}\left(X \leq x, Y_{c_0} \leq y, |X| > c_0\right) \\
&= \mathbf{P}\left(X \leq x, Y_{c_0} \leq y \mid |X| \leq c_0\right) \mathbf{P}\left(|X| \leq c_0\right) + \mathbf{P}\left(X \leq x, Y_{c_0} \leq y \mid |X| > c_0\right) \mathbf{P}\left(|X| > c_0\right) \\
&= \mathbf{P}\left(X \leq x, X \leq y \mid |X| \leq c_0\right) \mathbf{P}\left(|X| \leq c_0\right) + \mathbf{P}\left(X \leq x, -X \leq y \mid |X| > c_0\right) \mathbf{P}\left(|X| > c_0\right) \\
&= \mathbf{P}\left(X \leq x \wedge y \mid |X| \leq c_0\right) \mathbf{P}\left(|X| \leq c_0\right) + \mathbf{P}\left(-y \leq X \leq x \mid |X| > c_0\right) \mathbf{P}\left(|X| > c_0\right) \\
&= \mathbf{P}\left(X \leq x \wedge y, |X| \leq c_0\right) + \mathbf{P}\left(-y \leq X \leq x, |X| > c_0\right),
\end{aligned}$$

for all $x, y \in \mathbb{R}$. Therefore, we will have proved that X and Y_{c_0} are not independent once we show the existence of some $x_0, y_0 \in \mathbb{R}$ such that

$$\mathbf{P}\left(X \leq x_0\right) \mathbf{P}\left(Y_{c_0} \leq y_0\right) \neq \mathbf{P}\left(X \leq x_0 \wedge y_0, |X| \leq c_0\right) + \mathbf{P}\left(-y_0 \leq X \leq x_0, |X| > c_0\right).$$

For instance, consider $y_0 = x_0 = 0$. We have

$$\mathbf{P}\left(X \leq x_0\right) \mathbf{P}\left(Y_{c_0} \leq x_0\right) = \mathbf{P}\left(X \leq 0\right) \mathbf{P}\left(Y_{c_0} \leq 0\right) = \Phi^2\left(0\right) = \tfrac{1}{4}$$

and

$$\begin{aligned}
&\mathbf{P}\left(X \leq x_0 \wedge y_0, |X| \leq c_0\right) + \mathbf{P}\left(-y_0 \leq X \leq x_0, |X| > c_0\right) \\
s &= \mathbf{P}\left(X \leq 0, |X| \leq c_0\right) + \mathbf{P}\left(X = 0, |X| > c_0\right) \\
&= \mathbf{P}\left(-c_0 \leq X \leq 0\right) \\
&= \Phi\left(0\right) - \Phi\left(-c_0\right) \\
&= \tfrac{1}{2} - \Phi\left(-c_0\right)
\end{aligned}$$

Then, the above equalities would imply

$$\Phi\left(-c_0\right) = \tfrac{1}{4}.$$

On the other hand, it can computed that $c_0 \simeq 1.54$ and

$$\Phi\left(-1.54\right) \simeq 0.0618.$$

This shows that X and Y are not independent. □

Examples 754 and 755 show that, in general, a pair of uncorrelated random variables are not independent. Uncorrelation does not imply independence. However, there is an important exception of uncorrelated random variables which necessarily are independent.

Proposition 756 *Let X and Y be Bernoulli distributed, $X \sim Y \sim Ber\left(p\right)$, for some $p \in (0,1)$. Assume further X and Y are uncorrelated. Then, X and Y are independent.*

Proof. Since $X \sim Y \sim Ber(p)$, we have

$$\{X = 1, Y = 1\} = \{XY = 1\}$$

and

$$\{X = 1, Y = 0\} \cup \{X = 0, Y = 1\} \cup \{X = 0, Y = 0\} = \{XY = 0\}.$$

In addition,

$$\mathbf{E}[XY] = 1 \cdot \mathbf{P}(XY = 1) + 0 \cdot \mathbf{P}(XY = 0) = \mathbf{P}(XY = 1).$$

Under the assumption of the proposition, we have

$$\mathbf{P}(X = 1, Y = 1) = \mathbf{P}(XY = 1) = \mathbf{E}[XY] = \mathbf{E}[X]\,\mathbf{E}[Y] = \mathbf{P}(X = 1)\,\mathbf{P}(Y = 1). \quad (7.10)$$

and

$$\begin{aligned}
&\mathbf{P}(X = 1, Y = 0) + \mathbf{P}(X = 0, Y = 1) + \mathbf{P}(X = 0, Y = 0) \\
&= \mathbf{P}(XY = 0) \\
&= 1 - \mathbf{P}(XY = 1) \\
&= 1 - \mathbf{E}[XY] \\
&= 1 - \mathbf{E}[X]\,\mathbf{E}[Y] \\
&= 1 - \mathbf{P}(X = 1)\,\mathbf{P}(Y = 1). \qquad (7.11)
\end{aligned}$$

Now, since the events $\{X = 0, Y = 1\}$, $\{X = 0, Y = 0\}$ constitute a partition of $\{X = 0\}$, on account of (7.11), we can write

$$\begin{aligned}
&\mathbf{P}(X = 1, Y = 0) + \mathbf{P}(X = 0) \\
&= \mathbf{P}(X = 1, Y = 0) + \mathbf{P}(X = 0, Y = 1) + \mathbf{P}(X = 0, Y = 0) \\
&= 1 - \mathbf{P}(X = 1)\,\mathbf{P}(Y = 1) \\
&= 1 - \mathbf{P}(X = 1)(1 - \mathbf{P}(Y = 0)) \\
&= 1 - (\mathbf{P}(X = 1) - \mathbf{P}(X = 1)\,\mathbf{P}(Y = 0)) \\
&= 1 - \mathbf{P}(X = 1) + \mathbf{P}(X = 1)\,\mathbf{P}(Y = 0) \\
&= \mathbf{P}(X = 0) + \mathbf{P}(X = 1)\,\mathbf{P}(Y = 0)
\end{aligned}$$

which implies

$$\mathbf{P}(X = 1, Y = 0) = \mathbf{P}(X = 1)\,\mathbf{P}(Y = 0). \quad (7.12)$$

Similarly, since the events $\{X = 1, Y = 0\}$, $\{X = 0, Y = 0\}$ constitute a partition of $\{Y = 0\}$, still on account of (7.11), we can write

$$\begin{aligned}
&\mathbf{P}(X = 0, Y = 1) + \mathbf{P}(Y = 0) \\
&= \mathbf{P}(X = 0, Y = 1) + \mathbf{P}(X = 1, Y = 0) + \mathbf{P}(X = 0, Y = 0) \\
&= 1 - \mathbf{P}(X = 1)\,\mathbf{P}(Y = 1) \\
&= 1 - (1 - \mathbf{P}(X = 0))\,\mathbf{P}(Y = 1) \\
&= 1 - (\mathbf{P}(Y = 1) - \mathbf{P}(X = 0)\,\mathbf{P}(Y = 1)) \\
&= 1 - \mathbf{P}(Y = 1) + \mathbf{P}(X = 0)\,\mathbf{P}(Y = 1) \\
&= \mathbf{P}(Y = 0) + \mathbf{P}(X = 0)\,\mathbf{P}(Y = 1),
\end{aligned}$$

which implies

$$\mathbf{P}(X = 0, Y = 1) = \mathbf{P}(X = 0)\,\mathbf{P}(Y = 1). \quad (7.13)$$

In the end, thanks to (7.10), (7.12), and (7.13), we have

$$\begin{aligned}
\mathbf{P}(X=0,Y=0) &= 1-(\mathbf{P}(X=1,Y=1)+\mathbf{P}(X=1,Y=0)+\mathbf{P}(X=0,Y=1)) \\
&= 1-(\mathbf{P}(X=1)\mathbf{P}(Y=1)+\mathbf{P}(X=1)\mathbf{P}(Y=0)+\mathbf{P}(X=0)\mathbf{P}(Y=1)) \\
&= 1-(\mathbf{P}(X=1)(\mathbf{P}(Y=1)+\mathbf{P}(Y=0))+\mathbf{P}(X=0)\mathbf{P}(Y=1)) \\
&= 1-\mathbf{P}(X=1)-\mathbf{P}(X=0)\mathbf{P}(Y=1) \\
&= \mathbf{P}(X=0)-\mathbf{P}(X=0)\mathbf{P}(Y=1) \\
&= \mathbf{P}(X=0)(1-\mathbf{P}(Y=1)) \\
&= \mathbf{P}(X=0)\mathbf{P}(Y=0) \qquad (7.14)
\end{aligned}$$

Summing up, we have proved that

$$\begin{aligned}
&\mathbf{P}(X=1,Y=1)=\mathbf{P}(X=1)\mathbf{P}(Y=1), \quad \mathbf{P}(X=1,Y=0)=\mathbf{P}(X=1)\mathbf{P}(Y=0), \\
&\mathbf{P}(X=0,Y=1)=\mathbf{P}(X=0)\mathbf{P}(Y=1), \quad \mathbf{P}(X=0,Y=0)=\mathbf{P}(X=0)\mathbf{P}(Y=0),
\end{aligned}$$

(see Equations (7.10), (7.12), (7.13), and (7.14)). This clearly shows the independence of X and Y. □

Let $F_X : \mathbb{R} \to \mathbb{R}$ and $F_Y : \mathbb{R} \to \mathbb{R}$ be the distribution functions of the random variables X and Y, respectively, and let $F_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$ be the distribution function of the random vector (X, Y).

Proposition 757 *The random variables X and Y are independent if and only if we have*

$$F_{X,Y}(x,y) = F_X(x)\,F_Y(y), \qquad (7.15)$$

for every $(x, y) \in \mathbb{R}^2$.

Proof. The claim immediately follows from Proposition 750 by observing that

$$F_{X,Y}(x,y) = \mathbf{P}(X \in (-\infty, x], Y \in (-\infty, y])$$

and

$$F_X(x) = \mathbf{P}(X \in (-\infty, x]), \quad F_Y(y) = \mathbf{P}(Y \in (-\infty, y]).$$

□

Assume the random vector (X, Y) is absolutely continuous with density function $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}_+$ and let $f_X : \mathbb{R} \to \mathbb{R}$ and $f_Y : \mathbb{R} \to \mathbb{R}$ be the density functions of X and Y, respectively (see Proposition 714).

Proposition 758 *The random variables X and Y are independent if and only if we have*

$$f_{X,Y}(x,y) = f_X(x)\,f_Y(y),$$

μ_L^2-a.e. in $\mathbb{R}^2$.

Proof. Since the random vector (X,Y) is absolutely continuous with density function $f_{X,Y}:\mathbb{R}^2\to\mathbb{R}_+$, by definition we have

$$f_{X,Y}(x,y)=\frac{\partial^2 F_{X,Y}(x,y)}{\partial x\partial y},$$

μ_L^2-a.e. in $\mathbb{R}^2$. On the other hand, considering Equation (7.15) of Proposition 757, we can write

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

for every $(x,y)\in\mathbb{R}^2$. Hence,

$$f_{X,Y}(x,y)=\frac{\partial^2 F_X(x)\,F_Y(y)}{\partial x\partial y}=\frac{\partial F_X(x)}{\partial x}\frac{\partial F_Y(y)}{\partial y},\tag{7.16}$$

μ_L^2-a.e. in $\mathbb{R}^2$. Now, since the random variables X and Y are absolutely continuous (see Proposition 714) we have also

$$\frac{\partial F_X(x)}{\partial x}=f_X(x)\quad\text{and}\quad\frac{\partial F_Y(y)}{\partial y}=f_Y(y)\tag{7.17}$$

μ_L-a.e. in $\mathbb{R}$. Combining (7.16) and (7.17) the desired claim follows. □

Let $g:\mathbb{R}\to\mathbb{R}$ and $h:\mathbb{R}\to\mathbb{R}$ be Borel functions.

Proposition 759 *Assume X and Y are independent. Then, the real $\mathcal{E}$-random variables $g\circ X:\Omega\to\mathbb{R}$ and $h\circ Y:\Omega\to\mathbb{Y}$ are independent.*

Proof. . □

Example 760 *Referring to Example 754, we can state that the random variables X and X^2 are not independent because if they were independent, then, considering the Borel functions $g:\mathbb{R}\to\mathbb{R}$ and $h:\mathbb{R}\to\mathbb{R}$ given by*

$$g(x)\overset{def}{=}x^2\quad\text{and}\quad h(x)\overset{def}{=}x,\quad\forall x\in\mathbb{R},$$

also the random variables

$$g(X)=X^2\quad\text{and}\quad h\left(X^2\right)=X^2$$

should be independent. But, this is impossible since the only random variables which are independent of themselves are the Dirac random variables see Example (747).

Example 761 *Referring to Example 755, we can state that the random variables X and Y_c are not independent because if they were independent, then, considering the Borel functions $g:\mathbb{R}\to\mathbb{R}$ and $h:\mathbb{R}\to\mathbb{R}$ given by*

$$g(x)\overset{def}{=}x^2\quad\text{and}\quad h(x)\overset{def}{=}x^2,\quad\forall x\in\mathbb{R},$$

also the random variables

$$g(X)=X^2\quad\text{and}\quad h(Y_c)=X^2$$

should be independent. But, this is impossible (see Example 760).

Example 762 *Let X be a standard Gaussian random variable, $X \sim N(0,1)$, and let R be a standard Rademacher random variable, $R \sim Rad(1/2)$, which is independent of X. Consider the real random variable Y given by*

$$Y \overset{def}{=} RX$$

We have that $Y \sim N(0,1)$. In addition, R and Y are independent, while X and Y are uncorrelated, but not independent.

Discussion. Using the standard symbol Φ to denote the distribution function of $N(0,1)$, consider the distribution function F_Y of Y. Since the events $\{R=1\}$ and $\{R=-1\}$ constitute a partition of Ω, on account that X and R are independent and X is symmetric about 0, we can write

$$\begin{aligned}
\mathbf{P}(Y \leq y) &= \mathbf{P}(Y \leq y, R=1) + \mathbf{P}(Y \leq y, R=-1) \\
&= \mathbf{P}(RX \leq y \mid R=1)\,\mathbf{P}(R=1) + \mathbf{P}(RX \leq y \mid R=-1)\,\mathbf{P}(R=-1) \\
&= \tfrac{1}{2}\left(\mathbf{P}(X \leq y \mid R=1) + \mathbf{P}(-X \leq y \mid R=-1)\right) \\
&= \tfrac{1}{2}\left(\mathbf{P}(X \leq y) + \mathbf{P}(X \geq -y)\right) \\
&= \tfrac{1}{2}\left(\mathbf{P}(X \leq y) + \mathbf{P}(X \leq y)\right) \\
&= \mathbf{P}(X \leq y) \\
&= \Phi(X),
\end{aligned}$$

for every $y \in \mathbb{R}$. This proves that $Y \sim N(0,1)$. To prove the independence of R and Y, we compute

$$\mathbf{P}(R \leq r, Y \leq y) \quad \text{and} \quad \mathbf{P}(R \leq r)\,\mathbf{P}(Y \leq y),$$

on varying of $r, y \in \mathbb{R}$. We have

$$\mathbf{P}(R \leq r)\,\mathbf{P}(Y \leq y) = \begin{cases} 0, & \text{if } r < -1, \\ \frac{1}{2}\mathbf{P}(X \leq y), & \text{if } -1 \leq r < 1, \\ \mathbf{P}(X \leq y), & \text{if } 1 \leq r. \end{cases}$$

On the other hand,

$$\begin{aligned}
&\mathbf{P}(R \leq r, Y \leq y) \\
&= \mathbf{P}(R \leq r, Y \leq y, R=1) + \mathbf{P}(R \leq r, Y \leq y, R=-1) \\
&= \mathbf{P}(R \leq r, RX \leq y \mid R=1)\,\mathbf{P}(R=1) + \mathbf{P}(R \leq r, RX \leq y \mid R=-1)\,\mathbf{P}(R=-1) \\
&= \tfrac{1}{2}\left(\mathbf{P}(1 \leq r, X \leq y \mid R=1) + \mathbf{P}(-1 \leq r, -X \leq y \mid R=-1)\right) \\
&= \tfrac{1}{2}\left(\mathbf{P}(1 \leq r)\,\mathbf{P}(X \leq y) + \mathbf{P}(-1 \leq r)\,\mathbf{P}(-X \leq y)\right) \\
&= \tfrac{1}{2}\left(\mathbf{P}(1 \leq r)\,\mathbf{P}(X \leq y) + \mathbf{P}(-1 \leq r)\,\mathbf{P}(X \leq y)\right) \\
&= \tfrac{1}{2}\left(\mathbf{P}(1 \leq r) + \mathbf{P}(-1 \leq r)\right)\mathbf{P}(X \leq y),
\end{aligned}$$

where,

$$\mathbf{P}(1 \leq r) + \mathbf{P}(-1 \leq r) = \begin{cases} 0, & \text{if } r < -1, \\ 1, & \text{if } -1 \leq r < 1, \\ 2, & \text{if } 1 \leq r. \end{cases}$$

Therefore,

$$\mathbf{P}(R \leq r, Y \leq y) = \begin{cases} 0, & \text{if } r < -1, \\ \frac{1}{2}\mathbf{P}(X \leq y), & \text{if } -1 \leq r < 1, \\ \mathbf{P}(X \leq y), & \text{if } 1 \leq r. \end{cases}$$

It follows

$$\mathbf{P}(R \leq r, Y \leq y) = \mathbf{P}(R \leq r)\,\mathbf{P}(Y \leq y),$$

for all $r, y \in \mathbb{R}$, which is the desired result. Now, since R and X are independent we have that also R and X^2 are independent. Hence, on account of $\mathbf{E}[X] = 0$, it follows

$$\mathbf{E}[XY] = \mathbf{E}\left[RX^2\right] = \mathbf{E}[R]\,\mathbf{E}\left[X^2\right] = 0 = \mathbf{E}[X]\,\mathbf{E}[Y],$$

which shows that X and Y are uncorrelated. In the end, thanks to the independence of X and R, we have

$$\begin{aligned}
&\mathbf{P}(X \leq x, Y \leq y)\\
&= \mathbf{P}(X \leq x, Y \leq y, R = 1) + \mathbf{P}(X \leq x, Y \leq y, R = -1)\\
&= \mathbf{P}(X \leq x, RX \leq y \mid R = 1)\,\mathbf{P}(R = 1) + \mathbf{P}(X \leq x, RX \leq y \mid R = -1)\,\mathbf{P}(R = -1)\\
&= \tfrac{1}{2}\left(\mathbf{P}(X \leq x, X \leq y \mid R = 1) + \mathbf{P}(X \leq x, X \geq -y \mid R = 1)\right)\\
&= \tfrac{1}{2}\left(\mathbf{P}(X \leq x \wedge y \mid R = 1) + \mathbf{P}(-y \leq X \leq x \mid R = -1)\right)\\
&= \tfrac{1}{2}\left(\mathbf{P}(X \leq x \wedge y) + \mathbf{P}(-y \leq X \leq x)\right),
\end{aligned}$$

for all $x, y \in \mathbb{R}$. Therefore, we will prove that X and Y are not independent once we show the existence of some $x_0, y_0 \in \mathbb{R}$ such that

$$\mathbf{P}(X \leq x_0)\,\mathbf{P}(Y \leq y_0) \neq \tfrac{1}{2}\left(\mathbf{P}(X \leq x_0 \wedge y_0) + \mathbf{P}(-y_0 \leq X \leq x_0)\right).$$

For instance, consider $x_0 = -1, y_0 = 1$. We have

$$\mathbf{P}(X \leq x_0)\,\mathbf{P}(Y \leq y_0) = \mathbf{P}(X \leq -1)\,\mathbf{P}(Y \leq 1) = \Phi(-1)\,\Phi(1)$$

and

$$\tfrac{1}{2}\left(\mathbf{P}(X \leq x_0 \wedge y_0) + \mathbf{P}(-y_0 \leq X \leq x_0)\right) = \tfrac{1}{2}\left(\mathbf{P}(X \leq -1) + \mathbf{P}(-1 \leq X \leq -1)\right) = \tfrac{1}{2}\Phi(-1).$$

On the other hand,

$$\Phi(1) > \tfrac{1}{2} = \Phi(0).$$

As a consequence, X and Y are not independent.

On account of Proposition 759, to show that X and Y are not independent we could have argued as in Example 760. In fact, if X and Y were independent, considering the Borel functions $g : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$ given by

$$g(x) \overset{\text{def}}{=} x^2 \quad \text{and} \quad h(x) \overset{\text{def}}{=} x^2, \quad \forall x \in \mathbb{R},$$

also the random variables

$$g(X) = X^2 \quad \text{and} \quad h(Y) = Y^2 = R^2X^2 = X^2$$

should be independent. But, this is impossible (see Example 760). □

Let $\mathcal{F}$ a sub-σ-algebra of $\mathcal{E}$.

Definition 763 *We say that X is independent of $\mathcal{F}$ if $\sigma(X)$ is independent of $\mathcal{F}$, that is*

$$\mathbf{P}(E \cap F) = \mathbf{P}(E)\,\mathbf{P}(F), \tag{7.18}$$

for every $E \in \sigma(X)$ and every $F \in \mathcal{F}$. Equivalently,

$$\mathbf{P}(X \in B, F) = \mathbf{P}(X \in B)\,\mathbf{P}(F), \tag{7.19}$$

for every $B \in \mathcal{B}(\mathbb{R})$ and every $F \in \mathcal{F}$.

7.2 Family of Independent Real Random Variables

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space, let $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ the real Borel state space, let $(X_j)_{j\in J}$, where $X_j : \Omega \to \mathbb{R}$, for every $j \in J$, be a family of real $\mathcal{E}$-random variables on Ω, and let $\sigma(X_j)$ be the σ-algebra of events generated by the random variable X_j, for every $j \in J$.

Definition 764 *We say that the random variables of the family $(X_j)_{j\in J}$ are* pairwise independent *(with respect to $\mathbf{P}$), if for all $j_1, j_2 \in J$, such that $j_1 \neq j_2$, the random variables of the pair X_{j_1}, X_{j_2} are independent.*

Definition 765 *We say that the random variables of the family $(X_j)_{j\in J}$ are (*totally *or* mutually*)* independent *(with respect to $\mathbf{P}$), if the σ-algebras of the collection $(\sigma(X_j))_{j\in J}$ are (totally) independent.*

Proposition 766 *The random variables of the set $(X_j)_{j\in J}$ are independent if and only if we have*

$$\mathbf{P}(X_{j_1} \in B_1, \ldots, X_{j_n} \in B_n) = \mathbf{P}(X_{j_1} \in B_1) \cdot \cdots \cdot \mathbf{P}(X_{j_n} \in B_n), \tag{7.20}$$

for every finite subset $\{j_1, \ldots, j_n\}$ of J and for all $B_1, \ldots, B_n \in \mathcal{B}(\mathbb{R})$.

Proof. . □

Let $\mathcal{B}$ a basis for $\mathcal{B}(\mathbb{R})$.

Proposition 767 *The random variables of the family $(X_j)_{j\in J}$ are independent if and only if we have*

$$\mathbf{P}(X_{j_1} \in B_1, \ldots, X_{j_n} \in B_n) = \mathbf{P}(X_{j_1} \in B_1) \cdot \cdots \cdot \mathbf{P}(X_{j_n} \in B_n), \tag{7.21}$$

for every finite subset $\{j_1, \ldots, j_n\}$ of J and all $B_1, \ldots, B_n \in \mathcal{B}$.

Proof. . □

Let $P_{X_j} : \mathbb{R} \to \mathbb{R}$ be the distribution of the random variables X_j, for any $j \in J$, and let $P_{X_{j_1},\ldots,X_{j_n}} : \mathbb{R}^n \to \mathbb{R}_+$ the joint distribution of the random variables $X_{j_1}, \ldots, X_{j_n}$, for any finite subset $\{j_1, \ldots, j_n\}$ of J.

Proposition 768 *The random variables of the family $(X_j)_{j\in J}$ are independent if and only if we have*

$$P_{X_{j_1},\ldots,X_{j_n}}\left(\mathsf{X}_{j=1}^{n} B_j\right) = \prod_{k=1}^{n} P_{X_{j_1}}(B_j), \tag{7.22}$$

for any finite subset $\{j_1, \ldots, j_n\}$ of J and all $B_1, \ldots, B_n \in \mathcal{B}$.

Proof. . □

Proposition 769 *The random variables of the family $(X_j)_{j\in J}$ are independent if and only if we have*

$$P_{X_{j_1},\ldots,X_{j_n}} = P_{X_{j_1}} \otimes \cdots \otimes P_{X_{j_n}} \tag{7.23}$$

for any finite subset $\{j_1, \ldots, j_n\}$ of J.

Proof. . □

Proposition 770 *Assume the random variables of the family $(X_j)_{j\in J}$ are independent and $X_j \in \mathcal{L}^1(\Omega;\mathbb{R})$, for every $j \in J$. Then, the product random variable $X_{j_1}\cdots X_{j_n}$ ha finite moment of order 1, for any finite subset $\{j_1,\ldots,j_n\}$ of J, and we have*

$$\mathbf{E}[X_{j_1}\cdots X_{j_n}] = \mathbf{E}[X_{j_1}]\cdots\mathbf{E}[X_{j_n}]. \tag{7.24}$$

As a consequence, the random variables of the family $(X_j)_{j\in J}$ are pairwise uncorrelated, that is

$$Cov(X_{j_1}, X_{j_2}) = 0, \tag{7.25}$$

for every $j_1, j_2 \in J$.

Proof. .....Fixed any couple $j_1, j_2 \in J$ such that $j_1 \neq j_2$, since X_{j_1} are X_{j_2} independent, we have that $X_{j_1}X_{j_2} \in \mathcal{L}^1(\Omega;\mathbb{R})$ and

$$\mathbf{E}[X_{j_1}X_{j_2}] = \mathbf{E}[X_{j_1}]\mathbf{E}[X_{j_2}]. \tag{7.26}$$

Hence,

$$Cov(X_{j_1}, X_{j_2}) = \mathbf{E}[X_{j_1}X_{j_2}] - \mathbf{E}[X_{j_1}]\mathbf{E}[X_{j_2}] = 0, \tag{7.27}$$

as desired.. □

Let $F_{X_j} : \mathbb{R} \to \mathbb{R}$ be the distribution function of the random variables X_j, for any $j \in J$, and let $F_{X_{j_1},\ldots,X_{j_n}} : \mathbb{R}^n \to \mathbb{R}_+$ the joint distribution function of the random variables $X_{j_1},\ldots,X_{j_n}$, for any finite subset $\{j_1,\ldots,j_n\}$ of J.

Proposition 771 *The random variables of the family $(X_j)_{j\in J}$ are independent if and only if we have*

$$F_{X_{j_1},\ldots,X_{j_n}}(x_1,\ldots x_n) = F_{X_{j_1}}(x_1)\cdot\cdots\cdot F_{X_{j_n}}(x_n), \tag{7.28}$$

for every finite subset $\{j_1,\ldots,j_n\}$ of J and for every $(x_1,\ldots,x_n) \in \mathbb{R}$.

Proof. . □

Assume the real random variables of the family $(X_j)_{j\in J}$ are absolutely continuous. Let $f_{X_j} : \mathbb{R} \to \mathbb{R}$ be the density function of the random variables X_j, for any $j \in J$, and let $f_{X_{j_1},\ldots,X_{j_n}} : \mathbb{R}^n \to \mathbb{R}_+$ the joint density function of the random variables $X_{j_1},\ldots,X_{j_n}$, for any finite subset $\{j_1,\ldots,j_n\}$ of J.

Proposition 772 *The random variables of the set $(X_j)_{j\in J}$ are independent if and only if we have*

$$f_{X_{j_1},\ldots,X_{j_n}}(x_1,\ldots x_n) = f_{X_{j_1}}(x_1)\cdot\cdots\cdot f_{X_{j_n}}(x_n), \tag{7.29}$$

for every finite subset $\{j_1,\ldots,j_n\}$ of J and for μ_L^n-almost every $(x_1,\ldots x_n) \in \mathbb{R}^n$.

Proof. . □

Let $J_1,\ldots,J_m$ be pairwise disjoint finite subsets of J such that $J_h \equiv \{j_{h,1},\ldots j_{h,n_h}\}$ for every $h = 1,\ldots,m$, and let $g_h : \mathbb{R}^{n_h} \to \mathbb{R}$ be a Borel map, for every $h = 1,\ldots,m$.

Proposition 773 *Assume the random variables of the family $(X_j)_{j\in J}$ are independent. Then, the random variables $Y_1 : \Omega \to \mathbb{R},\ldots,Y_m : \Omega \to \mathbb{R}$ given by*

$$Y_h \overset{def}{=} g_h\left(X_{j_{h,1}},\ldots,X_{j_{h,n_h}}\right), \quad \forall h = 1,\ldots,m,$$

are independent.

Proof. . □

Example 774 *Let R_1 and R_2 two independent standard Rademacher random variables with success probability p, for some $p \in (0,1)$, and let Z be a standard Gaussian distributed random variable such that R_1, R_2, and Z are totally independent. Consider the random variables X_1 and X_2 given by*

$$X_k \overset{def}{=} R_k Z, \quad k = 1, 2.$$

We have that $X_1 \sim X_2 \sim N(0,1)$. In addition, X_1 and X_2 are uncorrelated, but not independent.

Discussion. Since R_k and Z are independent and the distribution of Z is symmetric about 0, we have

$$\begin{aligned}
\mathbf{P}(X_k \leq z) &= \mathbf{P}(R_k Z \leq z) \\
&= \mathbf{P}(R_k Z \leq z \mid R_k = 1)\,\mathbf{P}(R_k = 1) + \mathbf{P}(R_1 Z \leq z \mid R_k = -1)\,\mathbf{P}(R_k = -1) \\
&= \mathbf{P}(Z \leq z \mid R_k = 1)\,p + \mathbf{P}(Z \geq -z \mid R_k = -1)\,q \\
&= \mathbf{P}(Z \leq z)\,p + \mathbf{P}(Z \geq -z)\,q \\
&= \mathbf{P}(Z \leq z)\,(p+q) \\
&= \mathbf{P}(Z \leq z),
\end{aligned}$$

for $k = 1, 2$. This shows that $X_1 \sim X_2 \sim N(0,1)$. In addition, as a clear consequence, we have

$$\mathbf{P}(X_1 \leq x_1)\,\mathbf{P}(X_2 \leq x_2) = \mathbf{P}(Z \leq x_1)\,\mathbf{P}(Z \leq x_2).$$

On the other hand, thanks to the independence of R_1 and R_2 and the independence of Z from both R_1 and R_2, we have

$$\begin{aligned}
&\mathbf{P}(X_1 \leq x_1, X_2 \leq x_2) \\
&= \mathbf{P}(R_1 Z \leq x_1, R_2 Z \leq x_2 \mid R_1 = -1, R_2 = -1)\,\mathbf{P}(R_1 = -1, R_2 = -1) \\
&+ \mathbf{P}(R_1 Z \leq x_1, R_2 Z \leq x_2 \mid R_1 = -1, R_2 = 1)\,\mathbf{P}(R_1 = -1, R_2 = 1) \\
&+ \mathbf{P}(R_1 Z \leq x_1, R_2 Z \leq x_2 \mid R_1 = 1, R_2 = -1)\,\mathbf{P}(R_1 = 1, R_2 = -1) \\
&+ \mathbf{P}(R_1 Z \leq x_1, R_2 Z \leq x_2 \mid R_1 = 1, R_2 = 1)\,\mathbf{P}(R_1 = 1, R_2 = 1) \\
&= \mathbf{P}(-Z \leq x_1, -Z \leq x_2 \mid R_1 = -1, R_2 = -1)\,\mathbf{P}(R_1 = -1)\,\mathbf{P}(R_2 = -1) \\
&+ \mathbf{P}(-Z \leq x_1, Z \leq x_2 \mid R_1 = -1, R_2 = 1)\,\mathbf{P}(R_1 = -1)\,\mathbf{P}(R_2 = 1) \\
&+ \mathbf{P}(Z \leq x_1, -Z \leq x_2 \mid R_1 = 1, R_2 = -1)\,\mathbf{P}(R_1 = 1)\,\mathbf{P}(R_2 = -1) \\
&+ \mathbf{P}(Z \leq x_1, Z \leq x_2 \mid R_1 = 1, R_2 = -1)\,\mathbf{P}(R_1 = 1)\,\mathbf{P}(R_2 = 1) \\
&= \mathbf{P}(Z \geq -x_1, Z \geq -x_2)\,q^2 + \mathbf{P}(Z \geq -x_1, Z \leq x_2)\,pq \\
&+ \mathbf{P}(Z \leq x_1, Z \geq -x_2)\,pq + \mathbf{P}(Z \leq x_1, Z \leq x_2),
\end{aligned}$$

where $q \equiv 1 - p$. Now, choosing $x_1 \equiv 0$ and $x_2 \equiv 1$, we obtain

$$\begin{aligned}
\mathbf{P}(X_1 \leq x_1, X_2 \leq x_2) &= \mathbf{P}(Z \geq 0)\, q^2 + \mathbf{P}(0 \leq Z \leq 1)\, pq \\
&\quad + \mathbf{P}(-1 \leq Z \leq 0)\, pq + \mathbf{P}(Z \leq 0)\, p^2 \\
&= \frac{1}{2}\left(p^2 + q^2\right) + 2\mathbf{P}(0 \leq Z \leq 1)\, pq \\
&= \frac{1}{2}\left(p^2 + q^2\right) + 2\left(\mathbf{P}(Z \leq 1) - \mathbf{P}(Z \leq 0)\right) pq \\
&= \frac{1}{2}\left(p^2 + q^2\right) - pq + 2pq\mathbf{P}(Z \leq 1) \\
&= \frac{1}{2}(p - q)^2 + 2pq\mathbf{P}(Z \leq 1) \\
&= \frac{1}{2}(2p - 1)^2 + 2p(1 - p)\,\mathbf{P}(Z \leq 1) \\
&= \frac{1}{2}(2p - 1)^2 + 2p(1 - p) * 0.8413447.
\end{aligned}$$

On the other hand,

$$\mathbf{P}(X_1 \leq x_1)\,\mathbf{P}(X_2 \leq x_2) = \mathbf{P}(Z \leq 0)\,\mathbf{P}(Z \leq 1) = \frac{1}{2}\mathbf{P}(Z \leq 1) = 0.4206724$$

and

$$\frac{1}{2}(2p - 1)^2 + 2p(1 - p) * 0.8413447 = 0.420672,$$

for some $p \in (0, 1)$, if and only if $p = 1/2$. Hence, if $p \neq 1/2$, we have

$$\mathbf{P}(X_1 \leq 0, X_2 \leq 1) \neq \mathbf{P}(X_1 \leq 0)\,\mathbf{P}(X_2 \leq 1),$$

which shows that the random variables X_1 and X_2 cannot be independent. In case $p = 1/2$ we have

$$\mathbf{P}(X_1 \leq 0, X_2 \leq 1) = \mathbf{P}(X_1 \leq 0)\,\mathbf{P}(X_2 \leq 1),$$

but this is not enough to state that X_1 and X_2 are independent, not even uncorrelated. However, we consider

$$Cov(X_1, X_2) = \mathbf{E}[X_1 X_2] - \mathbf{E}[X_1]\,\mathbf{E}[X_2] = \mathbf{E}\left[R_1 R_2 Z^2\right] - \mathbf{E}[R_1 Z]\,\mathbf{E}[R_2 Z].$$

Since Z is independent of both R_1 and R_2, we have also that Z^2 is independent of $R_1 R_2$ (see Proposition 773). Therefore,

$$\mathbf{E}\left[R_1 R_2 Z^2\right] = \mathbf{E}[R_1]\,\mathbf{E}[R_2]\,\mathbf{E}\left[Z^2\right] = (p - q)^2 = (2p - 1)^2$$

and

$$\mathbf{E}[R_1 Z]\,\mathbf{E}[R_2 Z] = \mathbf{E}[R_1]\,\mathbf{E}[Z]\,\mathbf{E}[R_2]\,\mathbf{E}[Z] = 0.$$

Therefore, we obtain that X_1 and X_2 are correlated for every $p \in (0, 1)$ such that $p \neq 1/2$ and are uncorrelated for $p = 1/2$. We are still left with the task of showing whether the random variables X_1 and X_2 are independent for $p = 1/2$. To this, in light of Proposition 773, consider the random variables

$$Y_1 \stackrel{\text{def}}{=} X_1^2 \quad \text{and} \quad Y_1 \stackrel{\text{def}}{=} X_1^2.$$

We have

$$\mathbf{E}\left[Y_1Y_2\right] = \mathbf{E}\left[X_1^2X_2^2\right] = \mathbf{E}\left[R_1^2R_2^2Z^4\right] = \mathbf{E}\left[Z^4\right] = 3$$

and

$$\mathbf{E}\left[Y_1\right]\mathbf{E}\left[Y_2\right] = \mathbf{E}\left[X_1^2\right]\mathbf{E}\left[X_2^2\right] = \mathbf{E}\left[R_1^2Z^2\right]\mathbf{E}\left[R_2^2Z^2\right] = \mathbf{E}\left[Z^2\right]^2 = 1$$

It follows

$$\mathbf{E}\left[Y_1Y_2\right] - \mathbf{E}\left[Y_1\right]\mathbf{E}\left[Y_2\right] = 2.$$

Hence the random variables Y_1 and Y_2 are correlated, for every $p \in (0,1)$. This implies tha $\not$ Y_1 and Y_2 are not independent for every $p \in (0,1)$. Thus X_1 and X_2 cannot be independent for every $p \in (0,1)$. In particular, X_1 and X_2 are not independent for $p = 1/2$. □

Proposition 775 *Assume the random variables of the family $(X_j)_{j\in J}$ are pairwise independent, have finite moment of order 2, and have mean zero. Then, we have*

$$\mathbf{E}\left[\left(\sum_{h=1}^{n} X_{j_h}\right)^2\right] = \sum_{h=1}^{n} \mathbf{E}\left[X_{j_h}^2\right], \tag{7.30}$$

for every finite subset $\{j_1, \ldots, j_n\}$ of J.

Proof. Since the random variables of the family $(X_j)_{j\in J}$ are pairwise independent and have finite moment of order 2 the random variable $X_{j_1}X_{j_2}$ ha finite momentof order 1 for all $j_1, j_2 \in J$ such that $j_1 \neq j_2$. Hence, thanks to the identity

$$\left(\sum_{h=1}^{n} X_{j_h}\right)^2 = \sum_{h,k=1}^{n} X_{j_h}X_{j_k} = \sum_{h=1}^{n} X_{j_h}^2 + \sum_{\substack{h,k=1 \\ k\neq h}}^{n} X_{j_h}X_{j_k},$$

which holds true for every finite subset $\{j_1, \ldots, j_n\}$ of J, the linearity property of the expectation functional implies

$$\mathbf{E}\left[\left(\sum_{h=1}^{n} X_{j_h}\right)^2\right] = \sum_{h=1}^{n} \mathbf{E}\left[X_{j_h}^2\right] + \sum_{\substack{h,k=1 \\ k\neq h}}^{n} \mathbf{E}\left[X_{j_h}X_{j_k}\right].$$

Therefore, on account of Proposition 770 and the assumption of zero mean, the desired result immediately follows. □

Proposition 776 *Assume the random variables of the family $(X_j)_{j\in J}$ are pairwise independent and have finite moment of the second order. Then, we have*

$$\mathbf{D}^2\left[\sum_{h=1}^{n} X_{j_h}\right] = \sum_{h=1}^{n} \mathbf{D}^2\left[X_{j_h}\right], \tag{7.31}$$

for every finite subset $\{j_1, \ldots, j_n\}$ of J.

Proof. In fact, since the random variables have finite moment of order 2, the random variable $\sum_{h=1}^{n} X_{j_h}$ also has. Hence, we can write

$$\mathbf{D}^2\left[\sum_{h=1}^{n} X_{j_h}\right] = \mathbf{E}\left[\left(\sum_{h=1}^{n} X_{j_h}\right)^2\right] - \left(\mathbf{E}\left[\sum_{h=1}^{n} X_{j_h}\right]\right)^2, \tag{7.32}$$

where

$$\mathbf{E}\left[\left(\sum_{h=1}^{n} X_{j_h}\right)^2\right] = \sum_{h=1}^{n} \mathbf{E}\left[X_{j_h}^2\right] + \sum_{\substack{h,k=1 \\ k\neq h}}^{n} \mathbf{E}\left[X_{j_h} X_{j_k}\right] \tag{7.33}$$

and

$$\left(\mathbf{E}\left[\sum_{h=1}^{n} X_{j_h}\right]\right)^2 = \left(\sum_{h=1}^{n} \mathbf{E}\left[X_{j_h}\right]\right)^2 = \sum_{h=1}^{n} \mathbf{E}\left[X_{j_h}\right]^2 + \sum_{\substack{h,k=1 \\ k\neq h}}^{n} \mathbf{E}\left[X_{j_h}\right]\mathbf{E}\left[X_{j_k}\right]. \tag{7.34}$$

Combining (7.32)-(7.34), we obtain

$$\mathbf{D}^2\left[\sum_{h=1}^{n} X_{j_h}\right] = \sum_{h=1}^{n} \mathbf{E}\left[X_{j_h}^2\right] - \sum_{h=1}^{n} \mathbf{E}\left[X_{j_h}\right]^2 = \sum_{h=1}^{n} \left(\mathbf{E}\left[X_{j_h}^2\right] - \mathbf{E}\left[X_{j_h}\right]^2\right),$$

which is the desired result. □

Chapter 8

Characteristic Functions

8.1 Complex Random variables

Let $(\mathbb{C}, \mathcal{B}(\mathbb{C})) \equiv \mathbb{C}$ be the Euclidean complex plane equipped with the Borel σ-algebra $\mathcal{B}(\mathbb{C})$.

Recall 777 *We call* real part *[*imaginary part*] the real function* $\mathrm{Re} : \mathbb{C} \to \mathbb{R}$ *[resp.* $\mathrm{Im} : \mathbb{C} \to \mathbb{R}$*] given by*

$$\mathrm{Re}\,(z) \overset{def}{=} x, \quad [\mathrm{Im}\,(z) \overset{def}{=} y], \quad \forall z \in \mathbb{C},\ z \equiv x + iy,$$

where $x, y \in \mathbb{R}$ *and* i *is the* imaginary unit*.*

Recall 778 *We call* conjugate *the complex function* $\overline{\cdot} : \mathbb{C} \to \mathbb{C}$ *given by*

$$\overline{z} \overset{def}{=} x - iy, \quad \forall z \in \mathbb{C},\ z \equiv x + iy,$$

where $x, y \in \mathbb{R}$ *and* i *is the* imaginary unit*.*

Recall 779 *We call* modulus *the positive function* $|\cdot| : \mathbb{C} \to \mathbb{R}_+$ *given by*

$$|z| \overset{def}{=} \sqrt{x^2 + y^2}, \quad \forall z \in \mathbb{C},\ z \equiv x + iy,$$

where $x, y \in \mathbb{R}$ *and* i *is the* imaginary unit*.*

Recall 780 *We have*

$$\mathrm{Re}\,(z) = \frac{z + \overline{z}}{2}, \qquad \mathrm{Im}\,(z) = \frac{z - \overline{z}}{2i}, \qquad |z| = \sqrt{z\overline{z}},$$

for every $z \in \mathbb{C}$*.*

Recall 781 *We have*

$$\max\{|x|, |y|\} \leq |z| \leq |x| + |y|.$$

for every $z \in \mathbb{C}$*,* $z \equiv x + iy$*, where* $x, y \in \mathbb{R}$ *and* i *is the* imaginary unit*.*

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space and let $Z : \Omega \to \mathbb{C}$ be a complex function on Ω, denoted briefly by Z when no confusion can arise.

Definition 782 *As a particular case of Definition 1406 (see also Definition 402), we say that Z is a* complex $\mathcal{E}$-random variable *on Ω, if for every $B \in \mathcal{B}(\mathbb{C})$ the Z-inverse image of B is an event of the σ-algebra $\mathcal{E}$. In symbols,*

$$\{Z \in B\} \in \mathcal{E}, \quad \forall B \in \mathcal{B}(\mathbb{C}),$$

where we retain the notation $\{Z \in B\}$ as a shorthand for $\{\omega \in \Omega : Z(\omega) \in B\}$.

Notation 783 *We write $\mathcal{RV}(\Omega;\mathbb{C})$ the set of all complex $\mathcal{E}$-random variables on Ω. Formally,*

$$\mathcal{RV}(\Omega;\mathbb{C}) \equiv \{Z : \nleqslant \to \mathbb{C} : \{Z \in B\} \in \mathcal{E}, \quad \forall B \in \mathcal{B}(\mathbb{C})\}.$$

Remark 784 *The family $\sigma(Z)$ of elements of $\mathcal{P}(\Omega)$ given by*

$$\sigma(Z) \overset{def}{=} \{E \in \mathcal{P}(\Omega) \mid E = \{Z \in B\},\ B \in \mathcal{B}(\mathbb{C})\}$$

is a σ-algebra of events of Ω.

Definition 785 *We call $\sigma(Z)$ the σ-algebra of Ω generated by Z (see Definition **??**).*

Remark 786 *Z is a* complex random variable *on Ω if and only if*

$$\sigma(Z) \subseteq \mathcal{E}.$$

Definition 787 *We call* real part *[*imaginary part*] of Z the real function* $\mathrm{Re}(Z) : \Omega \to \mathbb{R}$ *[resp.* $\mathrm{Im}(Z) : \Omega \to \mathbb{R}$*] given by*

$$\mathrm{Re}(Z)(\omega) \overset{def}{=} \mathrm{Re}(Z(\omega)), \quad [(\mathrm{Im} \circ Z)(\omega) \overset{def}{=} \mathrm{Im}(Z(\omega))], \quad \forall \omega \in \Omega.$$

Setting $\mathrm{Re}(Z) \equiv X$ *and* $\mathrm{Im}(Z) \equiv Y$*, using a standard notation, we will also write $Z \equiv X + iY$.*

Proposition 788 *The complex function $Z \equiv X + iY$ is a random variable on Ω if and only if both the real part X and the imaginary part Y of Z are random variables.*

Proposition 789 *The set $\mathcal{RV}(\Omega;\mathbb{C})$ of all complex $\mathcal{E}$-random variables on Ω is a complex linear space.*

Definition 790 *We call* conjugate *of $Z \equiv X + iY$ the complex function $\bar{Z} : \Omega \to \mathbb{R}$ given by*

$$\bar{Z}(\omega) \overset{def}{=} \overline{Z(\omega)} = X(\omega) - iY(\omega), \quad \forall \omega \in \Omega.$$

Definition 791 *We call* modulus *of $Z \equiv X + iY$ the positive function $|Z| : \mathbb{C} \to \mathbb{R}_+$ given by*

$$|Z|(\omega) \overset{def}{=} |Z(\omega)| = \sqrt{X^2(\omega) + Y^2(\omega)}, \quad \forall \omega \in \Omega.$$

Remark 792 *We have*

$$\max\{|X|, |Y|\} \leq |Z| \leq |X| + |Y|.$$

Proof. The first inequality of the chain being evident, we only prove the second. To this, we just observe that

$$x^2 + y^2 \leq (|x| + |y|)^2 ,$$

for all $x, y \in \mathbb{R}$. Taking the square root of both the members of the latter, we Then, obtain

$$\sqrt{x^2 + y^2} \leq |x| + |y| ,$$

for all $x, y \in \mathbb{R}$. The desired result immediately follows. □

Remark 793 *If Z is a complex random variable on Ω, Then, the positive function $|Z|$ is a random variable on Ω. The converse is not true.*

Let $Z \in \mathcal{RV}(\Omega; \mathbb{C})$, $Z \equiv X + iY$.

Definition 794 *We say that Z has* finite expectation *if the positive random variable $|Z|$ has. If Z has finite expectation it is also said to have* finite first-order moment *or to be* integrable.

Remark 795 *The positive random variable $|Z|$ has finite expectation if and only if both the real random variables X and Y have.*

Let $Z \in \mathcal{RV}(\Omega; \mathbb{C})$, $Z \equiv X + iY$, having first order moment.

Definition 796 *We call* expectation *of Z the complex number*

$$\mathbf{E}[Z] \stackrel{def}{=} \mathbf{E}[X] + i\mathbf{E}[Y] .$$

We denote by $\mathcal{L}^1(\Omega; \mathbb{C})$ the set of all complex random variable on Ω having expectation.

Remark 797 *We have*

$$|\mathbf{E}[Z]| \leq \mathbf{E}[|Z|] .$$

Remark 798 *We have*

$$\mathbf{E}\left[\bar{Z}\right] = \overline{\mathbf{E}[Z]}$$

8.2 Characteristic Function of a Random Variable

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space, let $(\mathbb{R}, \mathcal{B}(\mathbb{R})) \equiv \mathbb{R}$ be the Euclidean real line equipped with the Borel σ-algebra, and let X be a real random variable on Ω.

Definition 799 *We call the function $\varphi_X : \mathbb{R} \to \mathbb{C}$ given by*

$$\varphi_X(t) \overset{def}{=} \mathbf{E}[e^{itX}] \equiv \int_\Omega e^{itX(\omega)} d\mathbf{P}, \quad \forall t \in \mathbb{R}.$$

the characteristic function *of X.*

Notation 800 *We will denote briefly φ_X the characteristic function of X, when no confusion can arise.*

Remark 801 *By virtue of Euler's formula, we have*

$$\left|e^{itX}\right| = \left|\cos(tX) + i\sin(tX)\right| = \cos^2(tX) + \sin^2(tX) = 1.$$

Hence, there exists

$$\mathbf{E}[\left|e^{itX}\right|] = 1.$$

As a consequence, the complex random variable e^{itX} has expectation given by

$$\mathbf{E}[e^{itX}] = \mathbf{E}[\cos(tX) + i\sin(tX)] = \mathbf{E}[\cos(tX)] + i\mathbf{E}[\sin(tX)] \in \mathbb{C},$$

for each $t \in \mathbb{R}$. Therefore, $\varphi_X(t)$ is well defined for every $t \in \mathbb{R}$. In particular,

$$\varphi_X(0) = 1.$$

Furthermore,

$$|\varphi_X(t)| = \left|\mathbf{E}[e^{itX}]\right| \leq \mathbf{E}[\left|e^{itX}\right|] = 1.$$

Lemma 802 *We have*

$$\left|e^{ix} - 1\right| \leq |x|, \quad \forall x \in \mathbb{R}. \tag{8.1}$$

More generally

$$\left|e^{ix} - e^{iy}\right| \leq |x - y|, \quad \forall x, y \in \mathbb{R}. \tag{8.2}$$

Proof. For every $x \in \mathbb{R}$, we have

$$\begin{aligned}\int_0^x e^{iu} du &= \int_0^x (\cos(u) + i\sin(u))\, du = \int_0^x \cos(u)\, du + i\int_0^x \sin(u)\, du \\ &= \sin(u)|_0^x - i\cos(u)|_0^x = \sin(x) - i(\cos(x) - 1),\end{aligned}$$

Therefore,

$$i\int_0^x e^{iu} du = i(\sin(x) - i(\cos(x) - 1)) = \cos(x) + i\sin(x) - 1 = e^{ix} - 1,$$

It Then, follows,

$$\left|e^{ix} - 1\right| = \left|i\int_0^x e^{iu} du\right| = \left|\int_0^x e^{iu} du\right| \leq \left|\int_0^x \left|e^{iu}\right| du\right| = \left|\int_0^x du\right| = |x|,$$

More generally,

$$\left|e^{ix} - e^{iy}\right| = \left|e^{iy}\left(e^{i(x-y)} - 1\right)\right| = \left|e^{iy}\right| \left|e^{i(x-y)} - 1\right| \leq |x - y|,$$

for all $x, y \in \mathbb{R}$. □

Let φ_X be the characteristic function of a real random variable X.

Proposition 803 *We have*

1. $\varphi_X(0) = 1$;
2. $|\varphi_X(t)| \leq 1$ *for every* $t \in \mathbb{R}$;
3. $\varphi_X(t)$ *is uniformly continuous;*
4. *if* X *has finite expectation, Then,* $\varphi_X(t)$ *is* $\mathbf{E}[|X|]$*-Lipschitz continuous.*

Proof. Properties 1 and 2 have already been presented in Remark **??**. With regard to 3, let $\varepsilon > 0$ and let K_ε such that

$$\mathbf{P}\left(|X| > K_\varepsilon\right) < \frac{\varepsilon}{3}{}^{1}.$$

On account of Inequality 8.2 of Lemma 802, we have

$$\begin{aligned}
|\varphi_X(t_1) - \varphi_X(t_2)| &= \left|\mathbf{E}[e^{it_1X}] - \mathbf{E}[e^{it_2X}]\right| \leq \mathbf{E}[\left|e^{it_1X} - e^{it_2X}\right|] = \int_\Omega \left|e^{it_1X(\omega)} - e^{it_2X(\omega)}\right| d\mathbf{P} \\
&\leq \int_{|X|\leq K_\varepsilon} \left|e^{it_1X(\omega)} - e^{it_2X(\omega)}\right| d\mathbf{P} + \int_{|X|>K_\varepsilon} \left|e^{it_1X(\omega)}\right| d\mathbf{P} + \int_{|X|>K_\varepsilon} \left|e^{it_1X(\omega)}\right| d\mathbf{P} \\
&\leq \int_{|x|\leq K_\varepsilon} \left|e^{it_1x} - e^{it_2x}\right| dP_X + 2\mathbf{P}\left(|X| > K_\varepsilon\right) \\
&\leq \int_{|x|\leq K_\varepsilon} |t_1 - t_2|\,|x|\, dP_X + \frac{2}{3}\varepsilon \\
&\leq |t_1 - t_2|\, K_\varepsilon + \frac{2}{3}\varepsilon.
\end{aligned}$$

Therefore, choosing t_1, t_2 such that $|t_1 - t_2| < \varepsilon/K_\varepsilon$, we obtain

$$|\varphi_X(t_1) - \varphi_X(t_2)| < \varepsilon.$$

This yields the uniform continuity of φ_X.

In the end, with regard to 4 we have

$$\begin{aligned}
|\varphi_X(t_1) - \varphi_X(t_2)| &= \left|\mathbf{E}[e^{it_1X}] - \mathbf{E}[e^{it_2X}]\right| = \left|\mathbf{E}[e^{it_1X} - e^{it_2X}]\right| \\
&= \left|\mathbf{E}[e^{it_2X}\left(e^{it_1X}e^{-it_2X} - 1\right)]\right| = \left|\mathbf{E}[e^{it_2X}\left(e^{i(t_1-t_2)X} - 1\right)]\right| \\
&\leq \mathbf{E}[\left|e^{it_2X}\left(e^{i(t_1-t_2)X} - 1\right)\right|] = \mathbf{E}[\left|e^{it_2X}\right|\left|e^{i(t_1-t_2)X} - 1\right|] \\
&= \mathbf{E}[\left|e^{i(t_1-t_2)X} - 1\right|] \leq \mathbf{E}[|(t_1 - t_2)\, X|] \\
&= |t_1 - t_2|\, \mathbf{E}[|X|],
\end{aligned}$$

for all $t_1, t_2 \in \mathbb{R}$. This is the desired result. □

Let P_X [resp. F_X] be the distribution [resp. distribution function] of X.

---

[1] Note that

$$\mathbf{P}\left(X > n\right) = 1 - \mathbf{P}\left(X \leq n\right)$$

and

$$\lim_{n\to\infty} \mathbf{P}\left(X \leq n\right) = \mathbf{P}\left(\lim_{n\to\infty} \{X \leq n\}\right) = \mathbf{P}\left(X \in \mathbb{R}\right) = 1.$$

Hence,

$$\lim_{n\to\infty} \mathbf{P}\left(X > n\right) = 0.$$

Remark 804 *We have*

$$\varphi_X(t) = \int_{\mathbb{R}} e^{itx} dP_X = \int_{\mathbb{R}} e^{itx} dF_X(x), \quad \forall t \in \mathbb{R}.$$

Remark 804 justifies why the characteristic function of X is also known as the Fourier transform of the probability distribution P_X or the distribution function F_X of X.

Remark 805 *Assume X is a discrete real random variable such that $X(\Omega) \equiv \{x_n\}_{n\in N}$ for some countable index set N, and write $\mathbf{P}\{X = x_n\} \equiv p_n$ for every $n \in N$. Then, we have*

$$\varphi_X(t) = \sum_{n\in N} e^{itx_n} p_n. \tag{8.3}$$

Remark 806 *Assume X is absolutely continuous and let f_X be the density of X. Then, we have*

$$\varphi_X(t) = \int_{\mathbb{R}} e^{itx} f_X(x) dx. \tag{8.4}$$

Example 807 *Let X be a Dirac random variable concentrated at x_0. Then, we have*

$$\varphi_X(t) = e^{itx_0}. \tag{8.5}$$

In particular, if $x_0 = 0$,

$$\varphi_X(t) = 1. \tag{8.6}$$

Example 808 *Let X be a standard Bernoulli random variable, $X \sim Ber(p)$. Then, we have*

$$\varphi_X(t) = pe^{it} + q. \tag{8.7}$$

Proof. According to the definition, $X(\Omega) \equiv \{0, 1\}$ and

$$\mathbf{P}\{X = 0\} = q, \qquad \mathbf{P}\{X = 1\} = p.$$

By (8.3) the desired (**??**) immediately follows. □

Example 809 *Let X be a standard Binomial random variable, $X \sim Bin(n, p)$. Then, we have*

$$\varphi_X(t) = (pe^{it} + q)^n. \tag{8.8}$$

Proof. According to the definition, $X(\Omega) \equiv \{0, 1, \ldots, n\}$ and

$$\mathbf{P}\{X = k\} = \binom{n}{k} p^k q^{n-k}, \quad \forall k = 0, \ldots, n.$$

Applying (8.3), it Then, follows

$$\varphi_X(t) = \sum_{k=0}^{n} e^{itk} \binom{n}{k} p^k q^{n-k} = \sum_{k=0}^{n} \binom{n}{k} (pe^{it})^k q^{n-k}, \quad \forall t \in \mathbb{R},$$

which is the desired (8.8). □

Example 810 *Let X be a standard Poisson random variable, $X \sim P(\lambda)$. Then, we have*

$$\varphi_X(t) = e^{\lambda(e^{it}-1)}. \tag{8.9}$$

Proof. According to the definition, $X(\Omega) \equiv \mathbb{Z}_+$ and

$$\mathbf{P}\{X = n\} = \frac{e^{-\lambda}\lambda^n}{n!}.$$

Applying (8.3), we obtain

$$\varphi_X(t) = \sum_{n=0}^{\infty} e^{itn}\frac{e^{-\lambda}\lambda^n}{n!} = e^{-\lambda}\sum_{n=0}^{\infty}\frac{(\lambda e^{it})^n}{n!} = e^{-\lambda}e^{\lambda e^{it}},$$

which is the desired (8.9). □

Example 811 *Let X be a uniform random variable, $X \sim Unif\,(a, b)$. Then, we have*

$$\varphi_X(t) = i\frac{e^{ita} - e^{itb}}{(b-a)t}.$$

In particular, if $X \sim U(-1, 1)$, we have

$$\varphi_X(t) = \frac{\sin(t)}{t}.$$

Proof. According to the definition

$$f_X(x) = \frac{1}{b-a}1_{[a,b]}(x).$$

Applying (8.4), it Then, follows

$$\varphi_X(t) = \int_{-\infty}^{+\infty} e^{itx}\frac{1}{b-a}1_{[a,b]}(x)\,dx = \frac{1}{b-a}\int_a^b e^{itx}\,dx = \frac{1}{(b-a)it}e^{itx}|_a^b = i\frac{e^{ita} - e^{itb}}{(b-a)t}.$$

In particular, if $X \sim U(-1, 1)$, we obtain

$$\varphi_X(t) = i\frac{e^{-it} - e^{it}}{2t} = \frac{e^{it} - e^{-it}}{2it} = \frac{\sin(t)}{t},$$

as desired. □

Example 812 *Let X be an exponential random variable, $X \sim Exp(\lambda)$. Then, we have*

$$\varphi_X(t) = \frac{\lambda}{\lambda - it}.$$

Proof. According to the definition

$$f_X(x) \stackrel{\text{def}}{=} \lambda e^{-\lambda x}1_{\mathbb{R}_+}(x), \quad \forall x \in \mathbb{R}.$$

Applying (8.4), we obtain

$$\begin{aligned}\varphi_X(t) &= \int_{-\infty}^{+\infty} e^{itx}\lambda e^{-\lambda x}1_{\mathbb{R}_+}(x)\,dx = \lambda\int_0^{+\infty} e^{-(\lambda-it)x}\,dx = \lambda\lim_{x\to+\infty}\int_0^x e^{-(\lambda-it)\xi}\,d\xi \\ &= -\frac{\lambda}{\lambda-it}\lim_{x\to+\infty} e^{-(\lambda-it)\xi}|_0^x = -\frac{\lambda}{\lambda-it}\lim_{x\to+\infty}\left(e^{-(\lambda-it)x} - 1\right) = \frac{\lambda}{\lambda-it},\end{aligned}$$

as desired. □

Example 813 *Let X be a gamma real random variable, $X \sim \Gamma(\alpha, \lambda)$ for $\alpha, \lambda > 0$. Then, we have*

$$\varphi_X(t) = \left(1 - i\frac{t}{\lambda}\right)^{-\alpha} = \left(\frac{\lambda}{\lambda - it}\right)^{\alpha}. \tag{8.10}$$

Proof. According to the definition

$$f_X(x) \stackrel{\text{def}}{=} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} 1_{\mathbb{R}_+}(x), \quad \forall x \in \mathbb{R}.$$

Recalling that

$$e^{itx} = \sum_{n=0}^{\infty} \frac{(itx)^n}{n!},$$

and that the series $\sum_{n=0}^{\infty} \frac{(itx)^n}{n!}$ is totally convergent, applying (8.4), we have

$$\begin{aligned}
\varphi_X(t) &= \int_{-\infty}^{+\infty} e^{itx} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} 1_{\mathbb{R}_+}(x)\, dx = \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^{+\infty} e^{itx} x^{\alpha-1} e^{-\lambda x}\, dx \\
&= \frac{\lambda^\alpha}{\Gamma(\alpha)} \lim_{y \to +\infty} \int_0^y \left(\sum_{n=0}^{\infty} \frac{i^n t^n}{n!} x^n \right) x^{\alpha-1} e^{-\lambda x}\, dx = \frac{\lambda^\alpha}{\Gamma(\alpha)} \sum_{n=0}^{\infty} \frac{i^n t^n}{n!} \lim_{y \to +\infty} \int_0^y x^{\alpha+n-1} e^{-\lambda x}\, dx \\
&= \frac{\lambda^\alpha}{\Gamma(\alpha)} \sum_{n=0}^{\infty} \frac{i^n t^n}{n!} \lambda^{-(\alpha+n)} \lim_{y \to +\infty} \int_0^y \lambda^{\alpha+n-1} x^{\alpha+n-1} e^{-\lambda x}\, d(\lambda x) \\
&= \frac{1}{\Gamma(\alpha)} \sum_{n=0}^{\infty} \frac{i^n}{n!} \left(\frac{t}{\lambda}\right)^n \lim_{y \to +\infty} \int_0^{\lambda y} \xi^{\alpha+n-1} e^{-\xi}\, d\xi \\
&= \frac{1}{\Gamma(\alpha)} \sum_{n=0}^{\infty} \frac{i^n}{n!} \left(\frac{t}{\lambda}\right)^n \int_0^{\infty} x^{\alpha+n-1} e^{-x}\, dx \\
&= \frac{1}{\Gamma(\alpha)} \sum_{n=0}^{\infty} \frac{i^n}{n!} \left(\frac{t}{\lambda}\right)^n \Gamma(\alpha+n),
\end{aligned}$$

Now, by virtue of the Property (**??**) of the Gamma distribution, we have

$$\Gamma(\alpha+n) = \alpha\,(\alpha+1)\cdots(\alpha+n-1)\,\Gamma(\alpha).$$

It Then, follows

$$\varphi_X(t) = \sum_{n=0}^{\infty} \frac{\alpha\,(\alpha+1)\cdots(\alpha+n-1)}{n!} \left(\frac{it}{\lambda}\right)^n.$$

On the other hand, by virtue of the properties of the negative binomial coefficient,

$$\begin{aligned}
\sum_{n=0}^{\infty} \frac{\alpha\,(\alpha+1)\cdots(\alpha+n-1)}{n!} \left(\frac{it}{\lambda}\right)^n &= \sum_{n=0}^{\infty} \binom{n+\alpha-1}{n} \left(\frac{it}{\lambda}\right)^n \\
&= \sum_{n=0}^{\infty} (-1)^n \binom{-\alpha}{n} \left(\frac{it}{\lambda}\right)^n \\
&= \sum_{n=0}^{\infty} \binom{-\alpha}{n} \left(-\frac{it}{\lambda}\right)^n \\
&= \left(1 - i\frac{t}{\lambda}\right)^{-\alpha},
\end{aligned}$$

which is the desired result. □

Let X, Y real random variables on Ω with distribution [resp. characteristic functions] P_X, P_Y [resp. φ_X, φ_Y], respectively.

Proposition 814 *Assume $Y = aX + b$. Then, we have*

$$\varphi_Y(t) = e^{ibt}\varphi_X(at), \quad \forall t \in \mathbb{R}. \tag{8.11}$$

In particular,

$$\varphi_{-X}(t) = \overline{\varphi_X(t)}, \quad \forall t \in \mathbb{R}. \tag{8.12}$$

Proof. By a straightforward computation,

$$\varphi_Y(t) = \mathbf{E}[e^{itY}] = \mathbf{E}[e^{it(aX+b)}] = e^{ibt}\mathbf{E}[e^{itaX}] = e^{ibt}\varphi_X(at).$$

In particular, if $a = -1$ and $b = 0$, we have

$$\varphi_{-X}(t) = \varphi_X(-t) = \mathbf{E}[e^{-itX}] = \overline{\mathbf{E}[e^{itX}]} = \overline{\varphi_X(t)},$$

as desired. □

Corollary 815 *Assume X is symmetric, Then, $\varphi_X(t)$ is a real-valued function.*

Proof. Since X is symmetric, $P_X = P_{-X}$. We Then, have

$$\varphi_{-X}(t) = \varphi_X(t).$$

On the other hand, by virtue of (8.12),

$$\varphi_{-X}(t) = \overline{\varphi_X(t)}.$$

It Then, follows,

$$\varphi_X(t) = \overline{\varphi_X(t)},$$

which yields the desired result. □

Theorem 816 *Assume $\varphi_X = \varphi_Y$. Then, $P_X = P_Y$.*

Proof. See Lukacs [**?**], (1964). □

Corollary 817 *Assume φ_X is a real-valued function, Then, X is symmetric.*

Proof. Under the assumption of the Corollary, on account of (8.12), we have

$$\varphi_{-X}(t) = \overline{\varphi_X(t)} = \varphi_X(t).$$

Hence, by virtue of Theorem 816, $P_X = P_{-X}$. This means that X is symmetric. □

Corollary 818 *A random variable X is symmetric if and only if the characteristic function φ_X is real-valued.*

Proof. Just recup Corollaries 815 and 817. □

Let $X_1, \ldots, X_n$ be real random variables on Ω and let $\varphi_{X_1}, \ldots, \varphi_{X_n}$ be the corresponding characteristic functions.

8.2.1 Characteristic Functions and Moments

Let X be a real random variable on Ω and let φ_X be the characteristic function of X.

Theorem 819 *Assume X has finite moments up to the order n included, where $n \in \mathbb{N}$. Then, φ_X is differentiable in $\mathbb{R}$ up to the order n included and for every $k = 1, \ldots, n$*

$$\varphi_X^{(k)}(t) = i^k \mathbf{E}[X^k e^{itX}], \quad \forall t \in \mathbb{R}. \tag{8.13}$$

In particular,

$$\varphi_X^{(k)}(0) = i^k \mathbf{E}[X^k]. \tag{8.14}$$

Proof. Thanks to the linearity of the expectation operator, we have

$$\begin{aligned}\varphi_X'(t) &\overset{\text{def}}{=} \lim_{\Delta t \to 0} \frac{\varphi_X(t+\Delta t) - \varphi_X(t)}{\Delta t} \overset{\text{def}}{=} \lim_{\Delta t \to 0} \frac{\mathbf{E}[e^{i(t+\Delta t)X}] - \mathbf{E}[e^{itX}]}{\Delta t} \\ &= \lim_{\Delta t \to 0} \mathbf{E}\left[\frac{e^{i(t+\Delta t)X} - e^{itX}}{\Delta t}\right] = \lim_{\Delta t \to 0} \mathbf{E}\left[\frac{e^{itX}\left(e^{i\Delta tX} - 1\right)}{\Delta t}\right],\end{aligned}$$

where

$$\left|\frac{e^{itX}\left(e^{i\Delta tX} - 1\right)}{\Delta t}\right| \leq \frac{|i\Delta tX|}{|\Delta t|} = |X|, \quad \forall \Delta t \in \mathbb{R} - \{0\}$$

and

$$\lim_{\Delta t \to 0} \frac{e^{itX}\left(e^{i\Delta tX} - 1\right)}{\Delta t} = iXe^{itX}.$$

Now, under the assumption of the theorem, the random variable $|X|$ is integrable. Hence, we can apply the Dominated Convergence Theorem obtaining

$$\lim_{\Delta t \to 0} \mathbf{E}\left[\frac{e^{itX}\left(e^{i\Delta tX} - 1\right)}{\Delta t}\right] = \mathbf{E}\left[\lim_{\Delta t \to 0} \frac{e^{itX}\left(e^{i\Delta tX} - 1\right)}{\Delta t}\right] = \mathbf{E}\left[iXe^{itX}\right] = i\mathbf{E}\left[Xe^{itX}\right]$$

This proves that φ_X is first order differentiable and

$$\varphi_X'(t) = i\mathbf{E}\left[Xe^{itX}\right], \quad \forall t \in \mathbb{R}.$$

In light of this result, let us assume that for some $k < n$ the statement of the theorem holds true and let us try to prove it for $k + 1 \leq n$. In fact, we have

$$\begin{aligned}\varphi_X^{(k+1)}(t) &\overset{\text{def}}{=} \lim_{\Delta t \to 0} \frac{\varphi_X^{(k)}(t+\Delta t) - \varphi_X^{(k)}(t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{i^k \mathbf{E}[X^k e^{i(t+\Delta t)X}] - i^k \mathbf{E}[X^k e^{itX}]}{\Delta t} \\ &= i^k \lim_{\Delta t \to 0} \mathbf{E}\left[X^k \frac{e^{i(t+\Delta t)X} - e^{itX}}{\Delta t}\right] = i^k \lim_{\Delta t \to 0} \mathbf{E}\left[X^k \frac{e^{itX}\left(e^{i\Delta tX} - 1\right)}{\Delta t}\right],\end{aligned}$$

where

$$\left|X^k \frac{e^{itX}\left(e^{i\Delta tX} - 1\right)}{\Delta t}\right| \leq \frac{\left|i\Delta tX^{k+1}\right|}{|\Delta t|} = \left|X^{k+1}\right|, \quad \forall \Delta t \in \mathbb{R} - \{0\}$$

and

$$\lim_{\Delta t \to 0} X^k \frac{e^{itX}\left(e^{i\Delta tX} - 1\right)}{\Delta t} = iX^{k+1}e^{itX}.$$

Again, under the assumption of the theorem, the random variable $\left|X^k\right|$ is integrable and we can apply the Dominated Convergence Theorem obtaining

$$\lim_{\Delta t\to 0}\mathbf{E}\left[X^k\frac{e^{itX}\left(e^{i\Delta tX}-1\right)}{\Delta t}\right]=\mathbf{E}\left[\lim_{\Delta t\to 0}X^k\frac{e^{itX}\left(e^{i\Delta tX}-1\right)}{\Delta t}\right]=\mathbf{E}\left[iX^{k+1}e^{itX}\right]=i\mathbf{E}\left[X^{k+1}e^{itX}\right].$$

This implies the differentiability of $\varphi_X^{(k)}$, that is the differentiabilty of φ_X up to the order $k+1$, and the equality

$$\varphi_X^{(k+1)}(t)=i^{k+1}\mathbf{E}\left[X^{k+1}e^{itX}\right],\quad \forall t\in\mathbb{R}.$$

Therefore, the Finite Induction Principle yields the desired result. □

Corollary 820 *Assume X has finite moments up to the order n included, where $n\in\mathbb{N}$. Then, we have*

$$\varphi_X(t)=\sum_{m=0}^{n}\frac{(it)^k\,\mathbf{E}[X^k]}{k!}+o(t^n),\quad \forall t\in\mathbb{R}.$$

Proof. By virtue of Theorem 819, φ_X is differentiable in $\mathbb{R}$ up to the order M included. Hence, the Taylor formula with starting point 0 yields

$$\varphi_X(t)=\sum_{n=0}^{n}\frac{\varphi_X^{(k)}(0)}{k!}t^k+o(t^n).$$

Therefore, on account of 8.14, the desired result immediately follows (see also Lukacs [**?**], 1964). □

Example 821 *Let X be a standard normal real random variable, $X\sim N(0,1)$. Then, we have*

$$\varphi_X(t)=e^{-\frac{t^2}{2}},\quad \forall t\in\mathbb{R}. \tag{8.15}$$

Proof. Since $X(\Omega)\equiv\mathbb{R}$ and X has a density given by $f_X(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$, applying (8.4), we have

$$\varphi_X(t)=\int_{-\infty}^{+\infty}e^{itx}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\,dx=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}e^{itx}e^{-\frac{x^2}{2}}\,dx.$$

Now, since X has finite first-order moment, we can apply Theorem **??** which yields the differentiability of φ_X and the equation

$$\varphi_X'(t)=i\mathbf{E}[Xe^{itX}]=\frac{1}{\sqrt{2\pi}}i\int_{-\infty}^{+\infty}xe^{itx}e^{-\frac{x^2}{2}}\,dx$$

(see (**??**)). Integrating by parts, we Then, obtain

$$\begin{aligned}\varphi_X'(t)&=-\frac{1}{\sqrt{2\pi}}i\int_{-\infty}^{+\infty}e^{itx}\,d\left(e^{-\frac{x^2}{2}}\right)=-\frac{1}{\sqrt{2\pi}}i\left(e^{itx}e^{-\frac{x^2}{2}}\Big|_{-\infty}^{+\infty}-\int_{-\infty}^{+\infty}e^{-\frac{x^2}{2}}\,d\left(e^{itx}\right)\right)\\&=\frac{1}{\sqrt{2\pi}}i\left(\int_{-\infty}^{+\infty}ite^{itx}e^{-\frac{x^2}{2}}\,dx\right)=-\frac{1}{\sqrt{2\pi}}t\int_{-\infty}^{+\infty}e^{itx}e^{-\frac{x^2}{2}}\,dx\\&=-t\varphi_X(t).\end{aligned}$$

Therefore, φ_X fulfills the linear differential equation

$$\varphi_X'(t)=-t\varphi_X(t),\quad \forall t\in\mathbb{R}. \tag{8.16}$$

Solving (8.16) and recalling that $\varphi_X(0)=1$ it Then, follows (8.15). □

Example 822 *Let X be a general normal real random variable, $X \sim N(\mu, \sigma^2)$. Then, we have*

$$\varphi_X(t) = e^{i\mu t} e^{-\frac{\sigma^2 t^2}{2}}, \quad \forall t \in \mathbb{R}. \tag{8.17}$$

Proof. We can write

$$X = \sigma Y + \mu$$

where $Y \sim N(\mu, \sigma)$. Therefore, combining (8.11) with (8.15), we obtain the desired (8.17). □

Theorem 823 *Assume φ_X is differentiable up to the order $2n$ included, where $n \in \mathbb{N}$. Then, X has finite moments up to the order $2n$ included and Equations (??) and (??) hold true.*

Theorem 824 *Assume φ_X is integrable. Then, X is absolutely continuous with density f_X given by*

$$f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-itx} \varphi_X(t)\, dt. \tag{8.18}$$

8.2.2 Characteristic Functions of Independent Random Variables

Proposition 825 *Assume $X_1, \ldots, X_n$ are independent. Then, we have*

$$\varphi_{X_1+\cdots+X_n} = \varphi_{X_1} \cdots \varphi_{X_n}.$$

Proof. By virtue of Proposition (??), the independence of the random variables $X_1, \ldots, X_n$ implies the independence of the random variables $e^{itX_1}, \ldots, e^{itX_n}$. Therefore, thanks to Proposition (??), we have

$$\varphi_X(t) = \mathbf{E}\left[e^{it(X_1+\cdots+X_n)}\right] = \mathbf{E}\left[e^{itX_1} \cdots e^{itX_n}\right] = \mathbf{E}\left[e^{itX_1}\right] \cdots \mathbf{E}\left[e^{itX_n}\right] = \varphi_{X_1}(t) \cdots \varphi_{X_n}(t),$$

as desired. □

Proposition 826 *Assume $X_k \backsim P(\lambda_k)$ is a Poisson random variable of parameter $\lambda_k > 0$, for $k = 1, \ldots, n$. Assume also that $X_1, \ldots, X_n$ are independent. Then, we have $X_1 + \cdots + X_n \backsim P(\lambda)$, where $\lambda \equiv \sum_{k=1}^n \lambda_k$.*

Proof. By virtue of Proposition 825 and Example **??**, we have

$$\varphi_{X_1+\cdots+X_n}(t) = \varphi_{X_1}(t) \cdots \varphi_{X_N}(t) = e^{\lambda_1(e^{it}-1)} \cdots e^{\lambda_n(e^{it}-1)} = e^{(\lambda_1+\cdots+\lambda_n)(e^{it}-1)}.$$

Therefore, thanks to Theorem **??**, the desired result follows. □

Proposition 827 *Assume $X_k \sim \Gamma(\alpha_k, \lambda)$ is a Gamma random variable on Ω of parameters $\alpha_k, \lambda > 0$, for $k = 1, \ldots, n$. Assume also that $X_1, \ldots, X_n$ are independent. Then, we have $X_1 + \cdots + X_n \backsim \Gamma(\alpha, \lambda)$, where $\alpha = \sum_{k=1}^n \alpha_k$.*

Proof. By virtue of Proposition 825 and Example 813, we have

$$\varphi_{X_1+\cdots+X_n}(t) = \varphi_{X_1}(t) \cdots \varphi_{Xn}(t) = \left(1 - i\frac{t}{\lambda}\right)^{-\alpha_1} \cdots \left(1 - i\frac{t}{\lambda}\right)^{-\alpha_n} = \left(1 - i\frac{t}{\lambda}\right)^{-\alpha}.$$

Therefore, thanks to Theorem **??**, we obtain the desired result. □

Let $X_k \sim N(\mu_k, \sigma_k^2)$ be a general normal random variable on Ω of parameters μ_k, σ_k, for $n = 1, \ldots, n$.

Proposition 828 *Assume $X_1, \dots, X_n$ are independent. Then, we have*

$$\sum_{n=1}^{n} X_k \sim N\left(\mu, \sigma^2\right).$$

where $\mu \equiv \sum_{k=1}^{n} \mu_k$ and $\sigma^2 \equiv \sum_{k=1}^{n} \sigma_k^2$.

Proof. By virtue of Proposition 825 and Example 822 we have

$$\begin{aligned}\varphi_X(t) &= \varphi_{X_1}(t) \cdots \varphi_{X_n}(t) = e^{i\mu_1 t} e^{-\frac{\sigma_1^2 t^2}{2}} \cdots e^{i\mu_n t} e^{-\frac{\sigma_n^2 t^2}{2}} \\ &= e^{i\mu_1 t} \cdots e^{i\mu_n t} e^{-\frac{\sigma_1^2 t^2}{2}} \cdots e^{-\frac{\sigma_n^2 t^2}{2}} = e^{i(\mu_1 + \cdots + \mu_n)t} e^{-\frac{\left(\sigma_1^2 + \cdots + \sigma_n^2\right)t^2}{2}}.\end{aligned}$$

Therefore, thanks to Theorem **??**, the desired result follows. □

8.2.3 Characteristic Functions of Real Random Vectors

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space, let $(\mathbb{R}^N, \mathcal{B}(\mathbb{R}^N)) \equiv \mathbb{R}^N$ be the real N-dimensional Euclidean space equipped with the Borel σ-algebra, for some $N \in \mathbb{N}$, and let $X \equiv (X_1, \dots, X_N)$ be an N-dimensional real random vector on Ω.

Definition 829 *We call the* characteristic function *of X the function $\varphi_X : \mathbb{R}^N \to \mathbb{C}$, briefly φ_X, given by*

$$\varphi_X(t) \stackrel{def}{=} \mathbf{E}[e^{i\langle t, X\rangle}] \equiv \int_\Omega e^{i\langle t, X(\omega)\rangle} d\mathbf{P} \equiv \int_\Omega e^{i\sum_{n=1}^{N} t_n X_n(\omega)} d\mathbf{P}, \quad \forall t \equiv (t_1, \dots, t_N) \in \mathbb{R}^N. \quad (8.19)$$

Proposition 830 *The characteristic function of X is well defined for every $t \equiv (t_1, \dots, t_N) \in \mathbb{R}^N$. In particular,*

$$\varphi_X(0) = 1.$$

Furthermore,

$$|\varphi_X(t)| = 1.$$

Proof. By virtue of Euler's formula, we have

$$\left|e^{i\langle t, X\rangle}\right| = \left|\cos\left(\langle t, X\rangle\right) + i \sin\left(\langle t, X\rangle\right)\right| = \cos^2\left(\langle t, X\rangle\right) + \sin^2\left(\langle t, X\rangle\right) = 1.$$

Hence, there exists

$$\mathbf{E}[\left|e^{i\langle t, X\rangle}\right|] = 1. \quad (8.20)$$

As a consequence, the complex random variable $e^{i\langle t, X\rangle} \equiv e^{i\left(\sum_{n=1}^{N} t_n X_n\right)}$ has expectation given by

$$\mathbf{E}[e^{i\langle t, X\rangle}] = \mathbf{E}[\cos\left(\langle t, X\rangle\right) + i \sin\left(\langle t, X\rangle\right)] = \mathbf{E}\left[\cos\left(\sum_{n=1}^{N} t_n X_n\right)\right] + i\, \mathbf{E}[\sin\left(\sum_{n=1}^{N} t_n X_n\right)] \in \mathbb{C},$$

for each $t \equiv (t_1, \dots, t_N) \in \mathbb{R}^N$. This means that $\varphi_X(t)$ is well defined. In particular, a trivial computation shows that

$$\varphi_X(0) = 1.$$

In the end, thanks to (8.20), we have

$$|\varphi_X(t)| = \left|\mathbf{E}[e^{i\langle t, X\rangle}]\right| \leq \mathbf{E}[\left|e^{i\langle t, X\rangle}\right|] = 1.$$

This completes the proof. □

Let φ_X be the characteristic function of X.

Remark 831 *The characteristic function φ_{X_n} of the nth entry X_n of X is given by*

$$\varphi_{X_n}(t) = \mathbf{E}[e^{i\langle \tilde{t}_n, X\rangle}], \quad \forall t \in \mathbb{R},$$

where $\tilde{t}_n \equiv (0, \dots 0, \underset{n}{t}, 0, \dots 0)$, for $n = 1, \dots, N$.

Let $P_X : \mathcal{B}\left(\mathbb{R}^N\right) \to \mathbb{R}_+$ be the distribution of X.

Remark 832 *We have*

$$\varphi_X(t) = \int_{\mathbb{R}^N} e^{i\langle t,x\rangle} dP_X = \int_{\mathbb{R}^N} e^{i\sum_{n=1}^N t_n x_n} dP_X, \quad \forall t \equiv (t_1, \dots, t_N) \in \mathbb{R}^N.$$

Remark 833 *Assume X is absolutely continuous and let $f_X : \mathbb{R}^N \to \mathbb{R}_+$ be the density of X. We have*

$$\varphi_X(t) = \int_{\mathbb{R}^N} e^{i\langle t,x\rangle} f_X(x)\, dx = \int_{\mathbb{R}^N} e^{i\sum_{n=1}^N t_n x_n} f_X(x_1, \dots, x_N)\, dx_1 \dots dx_N, \quad \forall t \equiv (t_1, \dots, t_N) \in \mathbb{R}^N.$$

Let $b \equiv (b_1, \dots, b_N) \in \mathbb{R}^N$ be an N-dimensional real vector, let $A \equiv (a_{j,n})_{j,n=1}^N \in \mathbb{R}^{N\times N}$ be an N-order matrix, let $Y \equiv (Y_1, \dots, Y_N)$ be the N-dimensional real random vector on Ω given by

$$Y \overset{\text{def}}{=} AX + b,$$

and let φ_Y be the characteristic function of Y.

Remark 834 *We have*

$$\varphi_Y(t) = e^{i\langle t,b\rangle}\varphi_X\left(A^\intercal t\right), \quad \forall t \equiv (t_1, \dots, t_N) \in \mathbb{R}^N,$$

where $A^\intercal$ is the transpose of A.

Proof. In fact, by virtue of the properties of the scalar product on $\mathbb{R}^N$, we have

$$\begin{aligned}\varphi_Y(t) &= \mathbf{E}[e^{i\langle t,Y\rangle}] = \mathbf{E}[e^{i\langle t,AX+b\rangle}] = \mathbf{E}[e^{i\langle t,AX\rangle + i\langle t,b\rangle}] \\ &= \mathbf{E}[e^{i\langle t,AX\rangle}e^{i\langle t,b\rangle}] = \mathbf{E}[e^{i\langle A^\intercal t,X\rangle}e^{i\langle t,b\rangle}] = e^{i\langle t,b\rangle}\mathbf{E}[e^{i\langle A^\intercal t,X\rangle}] \\ &= e^{i\langle t,b\rangle}\varphi_X\left(A^\intercal t\right),\end{aligned}$$

as desired. □

Theorem 835 *Assume X has finite moments up to the order M included, where $M \in \mathbb{N}$. Then, φ_X is differentiable in $\mathbb{R}^N$ up to the order N included and for every multiindex $\alpha \equiv (\alpha_1, \dots, \alpha_N)$ such that $|\alpha| \leq M$*

$$\frac{\partial^{|\alpha|}\varphi_X}{\partial t_1^{\alpha_1} \cdots \partial t_N^{\alpha_N}}(t) = i^{|\alpha|}\mathbf{E}[X_1^{\alpha_1} \cdots X_N^{\alpha_N} e^{i\langle t,X\rangle}], \quad \forall t \equiv (t_1, \dots, t_N) \in \mathbb{R}^{\mathbb{N}}. \tag{8.21}$$

In particular,

$$\frac{\partial \varphi_X}{\partial t_N}(0) = i\mathbf{E}[X_N], \quad \forall N = 1, \dots, N \tag{8.22}$$

and

$$\frac{\partial^2 \varphi_X}{\partial t_{N_1}\partial t_{N_1}}(0) = -\mathbf{E}[X_{N_1}X_{N_2}], \quad \forall N_1, N_2 = 1, \dots, N.$$

Let $X_1, \ldots, X_N$ real random variables on Ω and let $\varphi_{X_1}, \ldots, \varphi_{X_N}$ the corresponding characteristic functions. Consider the N-dimensional random vector $X \equiv (X_1, \ldots, X_N)$ and let φ_X be the characteristic function of X.

Proposition 836 *The random variables $X_1, \ldots, X_N$ are independent if and only if we have*

$$\varphi_X(t) = \varphi_{X_1}(t_1) \cdots \varphi_{X_N}(t_N), \quad \forall t \equiv (t_1, \ldots, t_N) \in \mathbb{R}^{\mathbb{N}}. \tag{8.23}$$

Proof. Since the function $\phi : \mathbb{R} \times \mathbb{R} \to \mathbb{C}$ given by

$$\phi(t, x) \stackrel{\text{def}}{=} e^{itx}, \quad \forall (t, x) \in \mathbb{R} \times \mathbb{R},$$

is continuous, the independence of the real random variables $X_1, \ldots, X_N$ implies the independence of the complex random variables $e^{itX_1}, \ldots, e^{itX_N}$. As a consequence, we have

$$\varphi_X(t) = \mathbf{E}[e^{i\langle t, X\rangle}] = \mathbf{E}\left[e^{it_1X_1} \cdots e^{it_NX_N}\right] = \mathbf{E}\left[e^{it_1X_1}\right] \cdots \mathbf{E}\left[e^{it_NX_N}\right] = \varphi_{X_1}(t_1) \cdots \varphi_{X_N}(t_N), \tag{8.24}$$

for every $t \equiv (t_1, \ldots, t_N) \in \mathbb{R}^{\mathbb{N}}$. Conversely, assume that (8.23) holds true. It is well known that it is possible to define real random variables $Y_1, \ldots, Y_N$ such that $Y_1, \ldots, Y_N$ are independent and Y_N has the same distribution of X_N for every $N = 1, \ldots, N$. Considering the random vector $Y \equiv (Y_1, \ldots, Y_N)$, by virtue of (8.24), we have

$$\varphi_Y(t) = \varphi_{Y_1}(t_1) \cdots \varphi_{Y_N}(t_N), \quad \forall t \equiv (t_1, \ldots, t_N) \in \mathbb{R}^{\mathbb{N}}. \tag{8.25}$$

On the other hand, Theorem 816 implies

$$\varphi_{Y_N} = \varphi_{X_N}, \quad \forall N = 1, \ldots, N. \tag{8.26}$$

Combining (8.25) with (8.26), we obtain

$$\varphi_Y(t) = \varphi_{X_1}(t_1) \cdots \varphi_{X_N}(t_N), \quad \forall t \equiv (t_1, \ldots, t_N) \in \mathbb{R}^{\mathbb{N}}.$$

Hence, our assumption yields

$$\varphi_Y = \varphi_X.$$

Applying again Theorem 816, it follows that the random vectors X and Y have the same distribution. Hence, the random variables $X_1, \ldots, X_N$ are independent, because they have the same joint distribution of the independent random variables $Y_1, \ldots, Y_N$. □

Chapter 9

Gaussian Random Vectors

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space and, for some $N \in \mathbb{N}$, let $X : \Omega \to \mathbb{R}^N$ be an $\mathcal{E}$-random vector with entries $X_1, \ldots, X_N$, with finite moments of order 2. In symbols, $X \in L^2\left(\Omega, \mathbb{R}^N\right)$.

Definition 837 *We say that the random vector X is* Gaussian *or* normally distributed, *if the random variable $\sum_{K=1}^N c_K X_K$ is Gaussian distributed for every $(c_1, \ldots, c_N) \in \mathbb{R}^N$. That is to say, for every $c \equiv (c_1, \ldots, c_N) \in \mathbb{R}^N$, there exist $\mu_c \in \mathbb{R}$ and $\sigma_c \in \mathbb{R}_+$ such that*

$$\sum_{K=1}^N c_K X_K \sim N\left(\mu_c, \sigma_c^2\right),$$

where by a normal distribution with $\sigma_c^2 = 0$ we mean the Dirac distribution concentrated at μ_c, for any $\mu_c \in \mathbb{R}$. In case the random vector $X \equiv (X_1, \ldots, X_N)^\intercal$ is Gaussian distributed, we also say that the random variables $X_1, \ldots, X_N$ are jointly Gaussian *or* normally distributed

Theorem 838 *The random vector $X \equiv (X_1, \ldots, X_N)^\intercal$ is Gaussian distributed if and only if there exist $\mu_X \equiv (\mu_1, \ldots, \mu_N) \in \mathbb{R}^N$, $A_X \in \mathbb{R}^{N \times M}$, and independent standard Gaussian random variables $Z_1, \ldots, Z_M$, for some $M \leq N$, such that*

$$(X_1, \ldots, X_N)^\intercal = \mu_X^\intercal + A_X \left(Z_1, \ldots, Z_M\right)^\intercal. \tag{9.1}$$

Proof. . □

Proposition 839 *If random vector $X \equiv (X_1, \ldots, X_N)^\intercal$ is Gaussian distributed, with reference to the notation of Theorem 838, we have*

$$\mu_X = \mathbf{E}\left[X\right],$$

that is

$$\mu_K = \mathbf{E}\left[X_K\right],$$

for every $K = 1, \ldots, N$, and

$$Var\left(X\right) = A_X A_X^\intercal \equiv \Sigma_X^2 \equiv \left(\sigma_{J,K}\right)_{J,K=1}^N,$$

where

$$\sigma_{J,K} \equiv Cov(X_J, X_K) = \mathbf{E}\left[\left(X_J - \mu_J\right)\left(X_K - \mu_K\right)\right],$$

for all $J, K = 1, \ldots, N$.

Proof. . □

Proposition 840 *Assume the random vector $X \equiv (X_1, \ldots, X_N)^\intercal$ is Gaussian distributed, let $b \equiv (b_1, \ldots, b_M)^\intercal$ an M-dimensional real vector, and let $A \equiv (a_{J,K})_{J=1,K=1}^{M,N}$ be an $M \times N$ full rank real matrix. Then, the M-dimensional real vector Y given by*

$$Y \overset{def}{=} b + AX$$

is Gaussian distributed.

Proof. . □

Definition 841 *We say that a Gaussian distributed random vector X is* non-degenerate *if the autocovariance matrix of X is positive definite.*

Proposition 842 *If the Gaussian random vector $X \equiv (X_1, \ldots, X_N)^\intercal$ is non-degenerate, then X is absolutely continuous with density $f_X : \mathbb{R}^N \to \mathbb{R}_+$ given by*

$$f_X(x_1, \ldots, x_N) = \frac{1}{\sqrt{(2\pi)^N \det(\Sigma^2)}} \exp\left(-\frac{1}{2}(x-\mu)^\intercal \left(\Sigma^2\right)^{-1}(x-\mu)\right), \tag{9.2}$$

for every $(x_1, \ldots, x_N) \in \mathbb{R}^N$, where

$$(x-\mu) \equiv (x_1 - \mu_1, \ldots, x_N - \mu_N)^\intercal, \quad \mu_K \equiv \mathbf{E}[X_K], \quad K = 1, \ldots, N$$

and

$$\Sigma^2 \equiv (\sigma_{J,K})_{J,K=1}^N, \quad \sigma_{J,K} \equiv \mathbf{E}[(X_J - \mu_J)(X_K - \mu_K)], \quad J, K = 1, \ldots, N.$$

Proof. . □

Example 843 *Let $X \equiv (X_1, X_2)$ a non-degenerate Gaussian vector, and let $f_X : \mathbb{R}^2 \to \mathbb{R}$ the density of X given by*

$$f_X(x_1, x_2) = \frac{1}{2\pi\sigma_{X_1}\sigma_{X_2}\sqrt{1-\rho_{X_1,X_2}^2}} \exp\left(-\frac{1}{2(1-\rho_{X_1,X_2}^2)}\left[\left(\frac{x_1-\mu_{X_1}}{\sigma_{X_1}}\right)^2 - 2\rho_{X_1,X_2}\left(\frac{x_1-\mu_{X_1}}{\sigma_{X_1}}\right)\left(\frac{x_2-\mu_{X_2}}{\sigma_{X_2}}\right) + \left(\frac{x_2-\mu_{X_2}}{\sigma_{X_2}}\right)^2\right]\right)$$

for every $(x_1, x_2) \in \mathbb{R}$, where $\mu_{X_1} \equiv \mathbf{E}[X_1]$, $\mu_{X_2} \equiv \mathbf{E}[X_2]$, $\sigma_{X_1}^2 \equiv \mathbf{D}^2[X_1]$, $\sigma_{X_2}^2 \equiv \mathbf{D}^2[X_2]$, and $\rho_{X_1,X_2} \equiv cov(X_1, X_2)/\sigma_{X_1}\sigma_{X_2}$. Then, the marginal densities $f_{X_1} : \mathbb{R} \to \mathbb{R}$ and $f_{X_2} : \mathbb{R} \to \mathbb{R}$ of the entries X_1 and X_2 of X, respectively, are given by

$$f_{X_1}(x_1) = \frac{1}{\sqrt{2\pi}\sigma_{X_1}} \exp\left(-\left(\frac{x_1-\mu_{X_1}}{\sigma_{X_1}}\right)^2\right), \tag{9.3}$$

for every $x_1 \in \mathbb{R}$, and

$$f_{X_2}(x_2) = \frac{1}{\sqrt{2\pi}\sigma_{X_2}} \exp\left(-\left(\frac{x_2-\mu_{X_2}}{\sigma_{X_2}}\right)^2\right), \tag{9.4}$$

for every $x_2 \in \mathbb{R}$, respectively.

Proof. A straightforward computation shows we can write

$$f_X(x_1,x_2)=\frac{1}{\sqrt{2\pi}\sigma_{X_1}}\exp\left(-\left(\frac{x_1-\mu_{X_1}}{\sigma_{X_1}}\right)^2\right)\frac{1}{\sqrt{2\pi}\sigma_{X_2}\sqrt{1-\rho^2_{X_1,X_2}}}\exp\left(-\frac{1}{2\sigma^2_{X_2}(1-\rho^2_{X_1,X_2})}\left(x_2-\mu_{X_2}-\rho_{X_1,X_2}\frac{\sigma_{X_2}}{\sigma_{X_1}}(x_1- \tag{9.5}$$

or

$$f_X(x_1,x_2)=\frac{1}{\sqrt{2\pi}\sigma_{X_2}}\exp\left(-\left(\frac{x_2-\mu_{X_2}}{\sigma_{X_2}}\right)^2\right)\frac{1}{\sqrt{2\pi}\sigma_{X_1}\sqrt{1-\rho^2_{X_1,X_2}}}\exp\left(-\frac{1}{2\sigma^2_{X_1}(1-\rho^2_{X_1,X_2})}\left(x_1-\mu_{X_1}-\rho_{X_1,X_2}\frac{\sigma_{X_1}}{\sigma_{X_2}}(x_2- \tag{9.6}$$

Hence, we have

$$\begin{aligned}
&f_{X_1}(x_1)\\
&=\int_{\mathbb{R}} f_{X_1,X_2}(x_1,x_2)d\mu_L(x_2)\\
&=\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}\sigma_{X_1}}\exp\left(-\left(\frac{x_1-\mu_{X_1}}{\sigma_{X_1}}\right)^2\right)\frac{1}{\sqrt{2\pi}\sigma_{X_2}\sqrt{1-\rho^2_{X_1,X_2}}}\exp\left(-\frac{1}{2\sigma^2_{X_2}(1-\rho^2_{X_1,X_2})}\left(x_2-\mu_{X_2}-\rho_{X_1,X_2}\frac{\sigma_{X_2}}{\sigma_{X_1}}(x_1-\mu_{X_1}\right.\right.\right.\\
&=\frac{1}{\sqrt{2\pi}\sigma_{X_1}}\exp\left(-\left(\frac{x_1-\mu_{X_1}}{\sigma_{X_1}}\right)^2\right)\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}\sigma_{X_2}\sqrt{1-\rho^2_{X_1,X_2}}}\exp\left(-\frac{1}{2\sigma^2_{X_2}(1-\rho^2_{X_1,X_2})}\left(x_2-\mu_{X_2}-\rho_{X_1,X_2}\frac{\sigma_{X_2}}{\sigma_{X_1}}(x_1-\mu\right.\right.\right.
\end{aligned} \tag{9.7}$$

and

$$\begin{aligned}
&f_{X_2}(x_2)\\
&=\int_{-\infty}^{+\infty} f_{X_1,X_2}(x_1,x_2)d\mu_L(x_1)\\
&=\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}\sigma_{X_2}}\exp\left(-\left(\frac{x_2-\mu_{X_2}}{\sigma_{X_2}}\right)^2\right)\frac{1}{\sqrt{2\pi}\sigma_{X_1}\sqrt{1-\rho^2_{X_1,X_2}}}\exp\left(-\frac{1}{2\sigma^2_{X_1}(1-\rho^2_{X_1,X_2})}\left(x_1-\mu_{X_1}-\rho_{X_1,X_2}\frac{\sigma_{X_1}}{\sigma_{X_2}}(x_2-\mu\right.\right.\right.\\
&=\frac{1}{\sqrt{2\pi}\sigma_{X_2}}\exp\left(-\left(\frac{x_2-\mu_{X_2}}{\sigma_{X_2}}\right)^2\right)\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}\sigma_{X_1}\sqrt{1-\rho^2_{X_1,X_2}}}\exp\left(-\frac{1}{2\sigma^2_{X_1}(1-\rho^2_{X_1,X_2})}\left(x_1-\mu_{X_1}-\rho_{X_1,X_2}\frac{\sigma_{X_1}}{\sigma_{X_2}}(x_2-\mu\right.\right.\right.
\end{aligned} \tag{9.8}$$

On the other hand, setting

$$\eta\equiv\frac{1}{\sqrt{2}\sigma_{X_2}\sqrt{1-\rho^2_{X_1,X_2}}}\left(x_2-\mu_{X_2}-\rho_{X_1,X_2}\frac{\sigma_{X_2}}{\sigma_{X_1}}(x_1-\mu_{X_1})\right),$$

we have

$$\begin{aligned}
&\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}\sigma_{X_2}\sqrt{1-\rho^2_{X_1,X_2}}}\exp\left(-\frac{1}{2\sigma^2_{X_2}(1-\rho^2_{X_1,X_2})}\left(x_2-\mu_{X_2}-\rho_{X_1,X_2}\frac{\sigma_{X_2}}{\sigma_{X_1}}(x_1-\mu_{X_1})\right)^2\right)dx_2\\
&=\frac{1}{\sqrt{\pi}}\int_{-\infty}^{+\infty}\exp\left(-\eta^2\right)d\eta\\
&=1,
\end{aligned} \tag{9.9}$$

and similarly, setting

$$\xi \equiv \frac{1}{\sqrt{2}\sigma_{X_1}\sqrt{1-\rho^2_{X_1,X_2}}}\left(x_1 - \mu_{X_1} - \rho_{X_1,X_2}\frac{\sigma_{X_1}}{\sigma_{X_2}}\left(x_2-\mu_{X_2}\right)\right),$$

we have

$$\begin{aligned}
&\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma_{X_1}\sqrt{1-\rho^2_{X_1,X_2}}} \exp\left(-\frac{1}{2\sigma^2_{X_1}(1-\rho^2_{X_1,X_2})}\left(x_1 - \mu_{X_1} - \rho_{X_1,X_2}\frac{\sigma_{X_1}}{\sigma_{X_2}}\left(x_2-\mu_{X_2}\right)\right)^2\right) dx_1 \\
&= \frac{1}{\sqrt{\pi}}\int_{-\infty}^{+\infty} \exp\left(-\xi^2\right) d\xi \\
&= 1. \qquad (9.10)
\end{aligned}$$

Therefore, (9.3) and (9.4) easily follow by combining (9.7) with (9.9) and (9.8) with (9.10), respectively. □

9.1 Independent and Normally Distributed Random Variables

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space and, for some $N \in \mathbb{N}$, let $X_1 : \Omega \to \mathbb{R}, \ldots, X_N : \Omega \to \mathbb{R}$ be real $\mathcal{E}$-random variables on Ω.

Theorem 844 *Assume $X_1, \ldots, X_N$ are pairwise uncorrelated and jointly Gaussian distributed. Then, $X_1, \ldots, X_N$ are independent.*

Proposition 845 *If $X_1, \ldots, X_N$ are pairwise uncorrelated and Gaussian distributed, then, in general, $X_1, \ldots, X_N$ are not jointly Gaussian distributed.*

Theorem 846 *Assume $X_1, \ldots, X_N$ are independent and Gaussian distributed. Then, $X_1, \ldots, X_N$ are jointly Gaussian distributed.*

Corollary 847 *If $X_1, \ldots, X_N$ are independent and Gaussian distributed, then, the linear combination*

$$\beta_0 + \sum_{K=1}^{N} \beta_K X_K$$

is a Gaussian random variable for every $\beta_0, \beta_1, \ldots, \beta_N \in \mathbb{R}$ such that $\sum\limits_{K=1}^{N} \beta_K^2 > 0$.

Proposition 848 *The random variables $X_1, \ldots, X_N$ are jointly Gaussian distributed if and only if there exist $\mu \equiv (\mu_1, \ldots, \mu_N) \in \mathbb{R}^N$ and $\Sigma \equiv (\sigma_{J,K})_{J,K=1}^N$ non-negative definite symmetric matrix such that*

$$\varphi_X(u_1, \ldots, u_N) \equiv \mathbf{E}\left[\exp\left(i\sum_{K=1}^N u_K X_K\right)\right] = \exp\left(-\frac{1}{2}\sum_{J,K=1}^N \sigma_{J,K} u_J u_K + i\sum_{K=1}^N \mu_K u_K\right), \qquad (9.11)$$

for every $(u_1, \ldots, u_N) \in \mathbb{R}^N$, where $\varphi_X : \mathbb{R}^N \to \mathbb{C}$ is the characteristic function of the random vector $X \equiv (X_1, \ldots, X_N)^\intercal$.

Chapter 10

Conditioning Random Variables

10.1 Conditional Expectation Given an Event

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space, let $\mathcal{L}^1(\Omega; \mathbb{R})$ be the linear space of all real random variables on Ω having finite moment of order one, and let F be an event of Ω such that $\mathbf{P}(F) > 0$.

Definition 849 *For any $X \in \mathcal{L}^1(\Omega; \mathbb{R})$ we call the real number*

$$\mathbf{E}[X \mid F] \overset{def}{=} \frac{1}{\mathbf{P}(F)} \int_F X d\mathbf{P} \tag{10.1}$$

the conditional expectation *of X given F.*

The conditional expectation of a random avriable X given an event F expresses the mean value taken by X given that F occurs weighted for the probability of $\mathbf{P}(F)$. In fact, we have

$$\int_F X d\mathbf{P} = \int_\Omega X 1_F d\mathbf{P} = \mathbf{E}[X 1_F].$$

Proposition 850 *Assume $X = 1_E$, where 1_E is the indicator function of some $E \in \mathcal{E}$. Then, we have*

$$\mathbf{E}[1_E \mid F] = \mathbf{P}(E \mid F). \tag{10.2}$$

Proof. We have

$$\mathbf{E}[1_E \mid F] = \frac{1}{\mathbf{P}(F)} \int_F 1_E d\mathbf{P} = \frac{1}{\mathbf{P}(F)} \int_{E \cap F} d\mathbf{P} = \frac{\mathbf{P}(E \cap F)}{\mathbf{P}(F)} = \mathbf{P}(E \mid F),$$

as desired. □

Proposition 850 shows how the notion of conditional probability can be easily derived from the notion of conditional expectation given an event. However, in some simple cases also the converse is true: the notion of conditional expectation given an event can be derived from the notion of conditional probability.

Proposition 851 *Assume $X \in \mathcal{L}^1(\Omega; \mathbb{R})$ is discrete. Write $X(\Omega) \equiv \{x_n\}_{n \in N}$, where $N \subseteq \mathbb{N}$. Then, we have*

$$\mathbf{E}[X \mid F] = \sum_{n \in N} x_n \mathbf{P}(E_n \mid F). \tag{10.3}$$

where $E_n \equiv \{X = x_n\}$, for every $n \in N$.

Proof. Since we can write

$$X = \sum_{n \in N} x_n 1_{E_n},$$

where $E_n \equiv \{X = x_n\}$, for every $n \in N$, thanks to the properties of the Lebesue integral and Equation 10.2, we have

$$\begin{aligned} \mathbf{E}[X \mid F] &= \frac{1}{\mathbf{P}(F)} \int_F \sum_{n \in N} x_n 1_{E_n} d\mathbf{P} = \sum_{n \in N} x_n \frac{1}{\mathbf{P}(F)} \int_F 1_{E_n} d\mathbf{P} \\ &= \sum_{n \in N} x_n \mathbf{E}[1_{E_n} \mid F] = \sum_{n \in N} x_n \mathbf{P}(E_n \mid F), \end{aligned}$$

as desired. □

Proposition 852 *Let $\mathbf{P}_F : \mathcal{E} \to \mathbb{R}_+$ be the conditional probability on Ω given F. Then, we have*

$$\mathbf{E}[X \mid F] = \int_\Omega X d\mathbf{P}_F, \tag{10.4}$$

for every $X \in \mathcal{L}^1(\Omega; \mathbb{R})$.

Proof. Let us consider first $X \equiv 1_E$, for some $E \in \mathcal{E}$. Thanks to Proposition 850, we have

$$\mathbf{E}[1_E \mid F] = \mathbf{P}(E \mid F) = \mathbf{P}_F(E) = \int_\Omega 1_E d\mathbf{P}_F.$$

Hence, (10.4) holds true for any indicator random variable. By a standard extension procedure, it is Then, possible to prove that (10.4) holds true for every $X \in \mathcal{L}^1(\Omega; \mathbb{R})$. □

10.2 Conditional Expectation Given a σ-Field of Events

10.2.1 Definitions and Basic Results in $\mathcal{L}^1(\Omega;\mathbb{R})$

Let $(\Omega,\mathcal{E},\mathbf{P})\equiv\Omega$ be a probability space. Let $\mathcal{F}$ be a sub-σ-algebra of $\mathcal{E}$, let $\mathbf{P}_{|\mathcal{F}}:\mathcal{F}\to\mathbb{R}$ be the restriction of the probability $\mathbf{P}$ to $\mathcal{F}$, given by

$$\mathbf{P}_{|\mathcal{F}}(F)\overset{\text{def}}{=}\mathbf{P}(F),\quad\forall F\in\mathcal{F},$$

and let $\Omega_{\mathcal{F}}\equiv(\Omega,\mathcal{F},\mathbf{P}_{|\mathcal{F}})$ the probability space with sample space Ω, σ-algebra of events $\mathcal{F}$, and probability $\mathbf{P}_{|\mathcal{F}}$. In the end, let $\mathcal{L}^1(\Omega;\mathbb{R})$ [resp. $\mathcal{L}^1(\Omega_{\mathcal{F}};\mathbb{R})$] be the semi-normed linear space of all real random variables on Ω [resp. on $\Omega_{\mathcal{F}}$] having finite moment of order one.

Remark 853 *For any $X\in\mathcal{L}^1(\Omega;\mathbb{R})$ the function $\mathbf{P}_{\mathcal{F}}^X:\mathcal{F}\to\mathbb{R}$, given by*

$$\mathbf{P}_{\mathcal{F}}^X(F)\overset{def}{=}\int_F X d\mathbf{P},\quad\forall F\in\mathcal{F},$$

is a real measure on $\Omega_{\mathcal{F}}$, with a bounded total variation[1]

$$\left\|\mathbf{P}_{\mathcal{F}}^X\right\|=\mathbf{E}[|X|],$$

which is absolutely continuous *with respect to $\mathbf{P}_{|\mathcal{F}}$.*

[1] Let $(\mathbb{X},\mathcal{M})\equiv\mathbb{X}$ be a measurable space.

Definition 854 *We call an extended real measure on $\mathbb{X}$ any map $\mu:\mathcal{M}\to\bar{\mathbb{R}}$ such that:*

1. *If there exists $M^\dagger\in\mathcal{M}$ such that $\mu(M^\dagger)=+\infty$ [resp. $\mu(M^\dagger)=-\infty$] we have*
$$\mu(M)>-\infty\ \text{[resp. } \mu(M)<+\infty]\quad\forall M\in\mathcal{M};$$
2. *$\mu(\varnothing)=0$;*
3. *$\mu(\bigcup_{n=1}^\infty M_n)=\sum_{n=1}^\infty\mu(M_n)$ for every sequence $(M_n)_{n\geq1}$ of disjoint elements in $\mathcal{M}$.*

Note that if $\mu:\mathcal{M}\to\bar{\mathbb{R}}$ is an extended real measure, then for every sequence $(M_n)_{n\geq1}$ of disjoint elements in $\mathcal{M}$ and for every bijection $\pi:\mathbb{N}\to\mathbb{N}$, we must have

$$\sum_{n=1}^\infty\mu(M_n)=\sum_{n=1}^\infty\mu(M_{\pi(n)}).$$

Let $\mu:\mathcal{M}\to\bar{\mathbb{R}}$ be an extended real measure on $\mathbb{X}$.

Theorem 855 (Hahn) *There exist $P,N\in\mathcal{M}$ such that:*

1. *$P\cup N=\mathbb{X}$ and $P\cap N=\varnothing$;*
2. *$\mu(E)\geq0$, for every $E\in\mathcal{M}$ such that $E\subseteq P$;*
3. *$\mu(E)\leq0$, for every $E\in\mathcal{M}$ such that $E\subseteq N$.*

Moreover, the decomposition (P,N) is essentially unique*, that is to say that for any other couple $\left(\tilde{P},\tilde{N}\right)$ fulfilling 1-3 the symmetric differences $P\Delta\tilde{P}$ and $N\Delta\tilde{N}$ are μ-*null sets in the strong sense*. Namely, we have $\mu(M)=0$ for every $M\in\mathcal{M}$ such that $M\subseteq P\Delta\tilde{P}$ or $M\subseteq N\Delta\tilde{N}$.*

Definition 856 *We call any couple (P,N) of measurable subsets of $\mathbb{X}$ fulfilling 1-3 of Theorem 855 a* Hahn decomposition *of $\mathbb{X}$.*

Theorem 863 (Radon-Nikodým) *For any $X \in \mathcal{L}^1(\Omega;\mathbb{R})$ there exists $X_{\mathcal{F}} \in \mathcal{L}^1(\Omega_{\mathcal{F}};\mathbb{R})$ such that*

$$\mathbf{P}_{\mathcal{F}}^{X}(F) = \int_F X_{\mathcal{F}} d\mathbf{P}_{|\mathcal{F}}, \tag{10.5}$$

for every $F \in \mathcal{F}$. Moreover, $X_{\mathcal{F}}$ is $\mathbf{P}_{|\mathcal{F}}$-almost surely uniquely determined. To say that if $Y_{\mathcal{F}} \in \mathcal{L}^1(\Omega_{\mathcal{F}};\mathbb{R})$ also satisfies Equation (10.5), then we have

$$Y_{\mathcal{F}} = X_{\mathcal{F}},$$

Theorem 857 (Jordan) *There exist $\mu^+ : \mathcal{M} \to \bar{\mathbb{R}}_+$ and $\mu^- : \mathcal{M} \to \bar{\mathbb{R}}_+$ positive measures on $\mathbb{X}$, at least one of which is finite, such that*

$$\mu = \mu^+ - \mu^-,$$

and for any Hahn decomposition (P,N) of $\mathbb{X}$ we have $\mu^+(M) = 0$ [resp. $\mu^-(M) = 0$], for every $M \in \mathcal{M}$ such that $M \subseteq N$ [resp. $M \subseteq P$].

Definition 858 *We call the couple (μ^+,μ^-) of positive measures introduced in Theorem 857 a* Jordan decomposition *of μ. In particular, we call μ^+ [resp. μ^-] the* positive [resp. negative] part *of μ.*

Proposition 859 *If (μ^+,μ^-) is a Jordan decomposition of μ we have*

$$\mu^+(M) = \mu(M \cap P), \quad \forall M \in \mathcal{M}$$

and

$$\mu^-(M) = -\mu(N \cap N), \quad \forall M \in \mathcal{M}$$

where (P,N) is any Hahn decomposition of $\mathbb{X}$.

Corollary 860 *If (μ^+,μ^-) is a Jordan decomposition of μ and both μ^+ and μ^- are finite we have*

$$\mu^+(M) = \sup_{G \in \mathcal{M}, G \subseteq M} \mu(G)$$

and

$$\mu^-(M) = -\inf_{G \in \mathcal{M}, G \subseteq E} \mu(G).$$

Moreover if $\nu^+ : \mathcal{M} \to \mathbb{R}_+$ and $\nu^- : \mathcal{M} \to \mathbb{R}_+$ are two finite real measures on $\mathbb{X}$ such that

$$\mu = \nu^+ - \nu^-,$$

we have

$$\nu^+ \geq \mu^+ \quad \text{and} \quad \nu^- \geq \mu^-.$$

Namely, the Jordan decomposition (μ^+,μ^-) of μ is the minimal decomposition of μ into a difference of two finite positive measures.

Let (μ^+,μ^-) be a Jordan decomposition of μ.

Definition 861 *We call* total variation *of μ the positive number*

$$\|\mu\| \overset{def}{=} \mu^+(\mathbb{X}) + \mu^-(\mathbb{X}).$$

We call total variation measure *of μ the positive measure $|\mu| : \mathcal{M} \to \bar{\mathbb{R}}_+$ given by*

$$|\mu| \overset{def}{=} \mu^+ + \mu^-.$$

Remark 862 *We clearly have*

$$\|\mu\| = |\mu|(\mathbb{X}).$$

$\mathbf{P}_{|\mathcal{F}}$*-a.s. on* $\Omega_{\mathcal{F}}$*. In the end, if* X *is* $\mathbf{P}$*-almost surely positive, then* $X_{\mathcal{F}}$ *is* $\mathbf{P}_{|\mathcal{F}}$*-almost surely positive.*

Definition 864 *We call a* version of the conditional expectation of X given $\mathcal{F}$ *any random variable in* $\mathcal{L}^1(\Omega_{\mathcal{F}};\mathbb{R})$ *satisfying Equation (10.5) of Theorem 863. We call the* conditional expectation of X given $\mathcal{F}$*, and we denote it by* $\mathbf{E}[X \mid \mathcal{F}]$*, the equivalence class of all the random variables in* $\mathcal{L}^1(\Omega_{\mathcal{F}};\mathbb{R})$ *satisfying (??) which are* $\mathbf{P}_{|\mathcal{F}}$*-almost surely equal.*

Likewise the former case of the density of a random variable, for our purposes we can neglect the distinction between the conditional expectation of a random variable given a σ-algebra of events, and the versions of such a conditional expectation. Therefore, by the conditional expectation of an event given a σ-algebra of events we will mean any version of the conditional expectation for which we use the notation $\mathbf{E}[X \mid \mathcal{F}]$ when no confusion can arise. In the particular case when the σ-algebra $\mathcal{F}$ is generated by an $\mathcal{E}$-random vector $Y : \Omega \to \mathbb{R}^N$, for some $N \in \mathbb{N}$, the conditional expectation of X given $\mathcal{F} \equiv \sigma(Y)$ is also called the *conditional expectation of* X *given* Y, and is more commonly denoted by $\mathbf{E}[X \mid Y]$ rather than $\mathbf{E}[X \mid \sigma(Y)]$.

Remark 865 *With the language of measure theory the conditional expectation of* X *given* $\mathcal{F}$ *is called the Radon-Nikodým derivative of* $\mathbf{P}^X_{\mathcal{F}}$ *with respect to* $\mathbf{P}_{|\mathcal{F}}$ *and is denoted by the symbol* $d\mathbf{P}^X_{\mathcal{F}}/d\mathbf{P}_{|\mathcal{F}}$*.*

Example 866 *With reference to Example 1415, let* $\Omega \equiv \{\omega_1, \ldots, \omega_6\}$ *be the sample space of all possible outcomes of the roll of a fair die. Consider the complete information* $\mathcal{E} \equiv \mathcal{P}(\Omega)$ *on* Ω*, the naive probability* $\mathbf{P} : \mathcal{E} \to \mathbb{R}_+$*, the state space* $(\mathbb{R}, \mathcal{B}(\mathbb{R})) \equiv \mathbb{R}$*, and let* $X : \Omega \to \mathbb{R}$ *the real* $\mathcal{E}$*-random variable given by*

$$X(\omega_j) \overset{def}{=} \begin{cases} j & \text{if } j \text{ is odd} \\ -j+1 & \text{if } j \text{ is even} \end{cases}, \quad \forall j = 1, \ldots, 6.$$

Setting $\mathcal{F} \equiv \{\varnothing, \Omega, F_{\mathbb{O}}, F_{\mathbb{E}}\}$*, where* $F_{\mathbb{O}} \equiv \{\omega_1, \omega_3, \omega_5\}$ *and* $F_{\mathbb{E}} \equiv \{\omega_2, \omega_4, \omega_6\}$*, we want to determine* $\mathbf{E}[X \mid \mathcal{F}]$*.*

Discussion. We already know that X is not an $\mathcal{F}$-random variable (see Example 1415). Nevertheless, $\mathbf{E}[X \mid \mathcal{F}]$ has to be a real $\mathcal{F}$-random variable. Thus, we will have

$$\mathbf{E}[X \mid \mathcal{F}] \neq X.$$

Since $\{F_{\mathbb{O}}, F_{\mathbb{E}}\}$ is a partition which generates $\mathcal{F}$, we can write

$$\mathbf{E}[X \mid \mathcal{F}] = x_1 1_{F_{\mathbb{O}}} + x_2 1_{F_{\mathbb{E}}},$$

where $x_1, x_2 \in \mathbb{R}$. Furthermore, we must have

$$\int_F \mathbf{E}[X \mid \mathcal{F}]\, d\mathbf{P}_{|\mathcal{F}} = \mathbf{P}^X_{\mathcal{F}}(F) = \int_F X d\mathbf{P}, \tag{10.6}$$

for every $F \in \mathcal{F}$. On the other hand,

$$\begin{aligned}\int_{F_{\mathbb{O}}} \mathbf{E}[X \mid \mathcal{F}]\, d\mathbf{P}_{|\mathcal{F}} &= \int_{F_{\mathbb{O}}} \left(x_1 1_{F_{\mathbb{O}}} + x_2 1_{F_{\mathbb{E}}}\right) d\mathbf{P}_{|\mathcal{F}} \\ &= x_1 \int_{F_{\mathbb{O}}} 1_{F_{\mathbb{O}}} d\mathbf{P}_{|\mathcal{F}} + x_2 \int_{F_{\mathbb{O}}} 1_{F_{\mathbb{E}}} d\mathbf{P}_{|\mathcal{F}} \\ &= x_1 \int_{F_{\mathbb{O}} \cap F_{\mathbb{O}}} d\mathbf{P}_{|\mathcal{F}} + x_2 \int_{F_{\mathbb{O}} \cap F_{\mathbb{E}}} d\mathbf{P}_{|\mathcal{F}} \\ &= x_1 \mathbf{P}_{|\mathcal{F}}(F_{\mathbb{O}}) + x_2 \mathbf{P}_{|\mathcal{F}}(\varnothing) \\ &= x_1 \mathbf{P}(F_{\mathbb{O}}) \\ &= \frac{1}{2} x_1.\end{aligned}$$

Similarly

$$\int_{F_{\mathbb{E}}} \mathbf{E}[X \mid \mathcal{F}]\, d\mathbf{P}_{|\mathcal{F}} = \frac{1}{2} x_2.$$

Therefore, Equation (10.6) implies

$$\frac{1}{2} x_1 = \int_{F_{\mathbb{O}}} X d\mathbf{P} \quad \text{and} \quad \frac{1}{2} x_2 = \int_{F_{\mathbb{E}}} X d\mathbf{P}.$$

We have

$$\int_{F_{\mathbb{O}}} X d\mathbf{P} = \int_{\Omega} X 1_{F_{\mathbb{O}}} d\mathbf{P} = \mathbf{E}\left[X 1_{F_{\mathbb{O}}}\right] = \sum_{j=1,3,5} j \mathbf{P}(X = j) = \frac{3}{2}$$

and

$$\int_{F_{\mathbb{E}}} X d\mathbf{P} = \int_{\Omega} X 1_{F_{\mathbb{E}}} d\mathbf{P} = \mathbf{E}[X 1_{F_{\mathbb{E}}}] = \sum_{j=2,4,3} (-j+1) \mathbf{P}(X = -j+1) = -\frac{3}{2}.$$

As a consequence,

$$\frac{1}{2} x_1 = \frac{3}{2} \quad \text{and} \quad \frac{1}{2} x_2 = -\frac{3}{2}.$$

In the end,

$$\mathbf{E}[X \mid \mathcal{F}] = 3 \cdot 1_{F_{\mathbb{O}}} - 3 \cdot 1_{F_{\mathbb{E}}}.$$

□

Example 867 *Consider $F \in \mathcal{E}$ such that $0 < \mathbf{P}(F) < 1$ and assume $\mathcal{F} = \sigma(1_F)$. Then, we have*

$$\mathbf{E}[X \mid \sigma(1_F)] \equiv \mathbf{E}[X \mid 1_F] = \mathbf{E}[X \mid 1_F = 1] 1_F + \mathbf{E}[X \mid 1_F = 0] 1_{F^c} \tag{10.7}$$

Discussion. We know that

$$\sigma(1_F) = \{\Omega, \varnothing, F, F^c\} \equiv \mathcal{F}.$$

Now, since $\mathbf{E}[X \mid 1_F]$ has to be a real $\mathcal{F}$-random variable, $\mathbf{E}[X \mid 1_F]$ has to be constant on both F and F^c. In addition, we must have

$$\int_F X d\mathbf{P} = \int_F \mathbf{E}[X \mid 1_F]\, d\mathbf{P} = \mathbf{E}[X \mid 1_F] \int_F d\mathbf{P} = \mathbf{E}[X \mid 1_F] \mathbf{P}(F)$$

and

$$\int_{F^c} X d\mathbf{P} = \int_{F^c} \mathbf{E}\left[X \mid 1_F\right] d\mathbf{P} = \mathbf{E}\left[X \mid 1_F\right] \int_{F^c} d\mathbf{P} = \mathbf{E}\left[X \mid 1_F\right] \mathbf{P}\left(F^c\right).$$

It then follows

$$\mathbf{E}\left[X \mid 1_F\right](\omega) = \begin{cases} \frac{1}{\mathbf{P}(F)} \int_F X d\mathbf{P}, & \text{if } \omega \in F, \\ \frac{1}{\mathbf{P}(F^c)} \int_{F^c} X d\mathbf{P}, & \text{if } \omega \in F^c, \end{cases}$$

which implies the desired result. □

The above example can be easily generalized as follows

Proposition 868 *Assume $\mathcal{F}$ is generated by a countable partition of events $(F_n)_{n\in N}$, where $N \subseteq \mathbb{N}$. Then, we have*

$$\mathbf{E}\left[X \mid \mathcal{F}\right] = \sum_{n\in N} \mathbf{E}\left[X \mid F_n\right] 1_{F_n}, \tag{10.8}$$

where 1_{F_n} is the indicator function of the event F_n, for every $n \in N$.

Proof. Since $\mathbf{E}\left[X \mid \mathcal{F}\right]$ has to be a real $\mathcal{F}$-random variable, under the assumption of the theorem, it follows $\mathbf{E}\left[X \mid \mathcal{F}\right]$ has to be constany on the event F_n, for every $n \in N$. Hence, etting

$$x_n \overset{\text{def}}{=} \mathbf{E}\left[X \mid \mathcal{F}\right](\omega), \quad \omega \in F_n, \quad \forall n \in N,$$

we can write

$$\mathbf{E}\left[X \mid \mathcal{F}\right] = \sum_{n\in N} x_n 1_{F_n}.$$

This implies

$$\begin{aligned} \mathbf{P}_{\mathcal{F}}^X(F_n) &= \int_{F_n} \mathbf{E}\left[X \mid \mathcal{F}\right] d\mathbf{P}_{|\mathcal{F}} = \int_{F_n} \left(\sum_{m\in N} x_m 1_{F_m}\right) d\mathbf{P}_{|\mathcal{F}} \\ &= \int_\Omega \left(\sum_{m\in N} x_m 1_{F_m}\right) 1_{F_n} d\mathbf{P}_{|\mathcal{F}} = \int_\Omega \sum_{\substack{m\in N \\ m\neq n}} x_m 1_{F_m\cap F_n} d\mathbf{P}_{|\mathcal{F}} + \int_\Omega x_n 1_{F_n} d\mathbf{P}_{|\mathcal{F}} \\ &= x_n \mathbf{P}\left(F_n\right), \end{aligned} \tag{10.9}$$

for every $n \in N$. On the other hand,

$$\mathbf{P}_{\mathcal{F}}^X(F_n) = \int_{F_n} X d\mathbf{P} = \mathbf{E}\left[X \mid F_n\right] \mathbf{P}\left(F_n\right), \tag{10.10}$$

Combining (10.9) and (10.10) if follows

$$x_n = \mathbf{E}\left[X \mid F_n\right],$$

for every $n \in N$, which is the desired result. □

10.2.2 Properties of the Conditional Expectation in $\mathcal{L}^1(\Omega;\mathbb{R})$

Let $\mathcal{L}^1(\Omega;\mathbb{R})$ [resp. $\mathcal{L}^1(\Omega_{\mathcal{F}};\mathbb{R})$] be the semi-normed linear space of all real random variables on Ω [resp. on $\Omega_{\mathcal{F}}$] having finite moment of order one.

Proposition 869 (invariance) *Assume $X \in \mathcal{L}^1(\Omega;\mathbb{R})$ itself is a real $\mathcal{F}$-random variable. Then*

$$\mathbf{E}[X \mid \mathcal{F}] = X.$$

In particular,

$$\mathbf{E}[\mathbf{E}[X \mid \mathcal{F}] \mid \mathcal{F}] = \mathbf{E}[X \mid \mathcal{F}].$$

Proposition 870 (mean preserving) *For any $X \in \mathcal{L}^1(\Omega;\mathbb{R})$ we have*

$$\int_F X d\mathbf{P} = \int_F \mathbf{E}[X \mid \mathcal{F}] \, d\mathbf{P}_{|\mathcal{F}}, \tag{10.11}$$

or, equivalently,

$$\mathbf{E}[X1_F] = \mathbf{E}[\mathbf{E}[X \mid \mathcal{F}] 1_F], \tag{10.12}$$

for every $F \in \mathcal{F}$. In particular,

$$\mathbf{E}[X] = \mathbf{E}[\mathbf{E}[X \mid \mathcal{F}]]. \tag{10.13}$$

Proposition 871 (linearity) *We have*

$$\mathbf{E}[\alpha X + \beta Y \mid \mathcal{F}] = \alpha \mathbf{E}[X \mid \mathcal{F}] + \beta \mathbf{E}[Y \mid \mathcal{F}], \tag{10.14}$$

for all $X, Y \in \mathcal{L}^1(\Omega;\mathbb{R})$ and all $\alpha, \beta \in \mathbb{R}$.

Proposition 872 (monotonicity) *We have*

$$\mathbf{E}[X \mid \mathcal{F}] \leq \mathbf{E}[Y \mid \mathcal{F}],$$

for all $X, Y \in \mathcal{L}^1(\Omega;\mathbb{R})$ such that $X \leq Y$.

Proposition 873 (law of iterated expectations - tower property) *Let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{E}$ such that $\mathcal{G} \subseteq \mathcal{F}$. Then*

$$\mathbf{E}[\mathbf{E}[X \mid \mathcal{F}] \mid \mathcal{G}] = \mathbf{E}[\mathbf{E}[X \mid \mathcal{G}] \mid \mathcal{F}] = \mathbf{E}[X \mid \mathcal{G}]. \tag{10.15}$$

for every $X \in \mathcal{L}^1(\Omega;\mathbb{R})$.

Proof. see Chung (1974) [**?**, thm 9.1.4, p. 300] □

Proposition 874 (concentration) *Assume $X \in \mathcal{L}^1(\Omega;\mathbb{R})$ is independent of $\mathcal{F}$. Then*

$$\mathbf{E}[X \mid \mathcal{F}] = \mathbf{E}[X]. \tag{10.16}$$

Proof. The deterministic random variable $\mathbf{E}[X]$ clearly belongs to $\mathcal{L}^1(\Omega_{\mathcal{F}};\mathbb{R})$. Under the independence assumption, we can write

$$\int_F X d\mathbf{P} = \int_\Omega X1_F d\mathbf{P} = \mathbf{E}[X1_F] = \mathbf{E}[X]\,\mathbf{E}[1_F] = \mathbf{E}[X] \int_\Omega 1_F d\mathbf{P} = \int_\Omega \mathbf{E}[X]\, 1_F d\mathbf{P} = \int_F \mathbf{E}[X] \, d\mathbf{P},$$

for every $F \in \mathcal{F}$. Therefore, $\mathbf{E}[X]$ satisfies Equation (10.5). This implies the desired result.

Proposition 875 (irrelevance of independent information) *Let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{E}$ which is independent of $\mathcal{F}$. Then, for any $X \in \mathcal{L}^1(\Omega;\mathbb{R})$ which is independent of $\mathcal{G}$ we have*

$$\mathbf{E}[X \mid \mathcal{F} \vee \mathcal{G}] = \mathbf{E}[X \mid \mathcal{F}].$$

Theorem 876 (transparence property) *We have*

$$\mathbf{E}[XY \mid \mathcal{F}] = Y\mathbf{E}[X \mid \mathcal{F}]. \tag{10.17}$$

For all $X, Y \in \mathcal{L}^1(\Omega;\mathbb{R})$ such that $XY \in \mathcal{L}^1(\Omega;\mathbb{R})$ and $Y \in \mathcal{L}^1(\Omega_{\mathcal{F}};\mathbb{R})$.

Theorem 877 *Given any convex function $\phi : \mathbb{R} \to \mathbb{R}$ such that $\phi \circ X \in \mathcal{L}^1(\Omega;\mathbb{R})$ we have*

$$\phi(\mathbf{E}[X \mid \mathcal{F}]) \leq \mathbf{E}[\phi(X) \mid \mathcal{F}], \tag{10.18}$$

for every $X \in \mathcal{L}^1(\Omega;\mathbb{R})$.

Proof. see Chung (1974) [**?**] □

Corollary 878 *We have*

$$|\mathbf{E}[X \mid \mathcal{F}]| \leq \mathbf{E}[|X| \mid \mathcal{F}],$$

for every $X \in \mathcal{L}^1(\Omega;\mathbb{R})$.

Proposition 879 *The conditional expectation operator $\mathbf{E}[\cdot \mid \mathcal{F}] : L^1(\Omega;\mathbb{R}) \to L^1(\Omega_{\mathcal{F}};\mathbb{R})$ is a positive linear operator of unit norm, which is a projection of $L^1(\Omega;\mathbb{R})$ onto $L^1(\Omega_{\mathcal{F}};\mathbb{R})$.*

Proof. By virtue of Theorem 863 the operator $\mathbf{E}[\cdot \mid \mathcal{F}] : L^1(\Omega;\mathbb{R}) \to L^1(\Omega_{\mathcal{F}};\mathbb{R})$ is well defined and positive, by virtue of Proposition 871 it is linear, and by virtue of Proposition 869, it is a map of $L^1(\Omega;\mathbb{R})$ onto $L^1(\Omega_{\mathcal{F}};\mathbb{R})$. We are left with showining that it has unit norn. To this, recall that

$$\|\mathbf{E}[\cdot \mid \mathcal{F}]\| = \sup\left\{\frac{\|\mathbf{E}[X \mid \mathcal{F}]\|_1}{\|X\|_1}, \quad X \in L^1(\Omega;\mathbb{R}) - \{0\}\right\}.$$

Now, we have

$$\mathbf{E}[X \mid \mathcal{F}] = X$$

for every $X \in L^1(\Omega_{\mathcal{F}};\mathbb{R})$. This clearly implies that

$$\frac{\|\mathbf{E}[X \mid \mathcal{F}]\|_1}{\|X\|_1} = 1 \tag{10.19}$$

for every $X \in L^1(\Omega_{\mathcal{F}};\mathbb{R}) - \{0\}$. Hence,

$$\|\mathbf{E}[\cdot \mid \mathcal{F}]\| \geq 1.$$

On the other hand, considering Corollary 878, we have

$$\|\mathbf{E}[X \mid \mathcal{F}]\|_1 = \int_\Omega |\mathbf{E}[X \mid \mathcal{F}]|\, d\mathbf{P}_{|\mathcal{F}} \leq \int_\Omega \mathbf{E}[|X| \mid \mathcal{F}]\, d\mathbf{P}_{|\mathcal{F}} = \int_\Omega |X|\, d\mathbf{P} = \|X\|_1.$$

It then follows

$$\frac{\|\mathbf{E}[X \mid \mathcal{F}]\|_1}{\|X\|_1} \leq 1, \tag{10.20}$$

for every $X \in L^1(\Omega;\mathbb{R}) - \{0\}$. Combining Equtions (10.19) and (10.20) we obtain thaty

$$\|\mathbf{E}[\cdot \mid \mathcal{F}]\| = 1,$$

as desired. □

10.2.3 Properties of the Conditional Expectation in $\mathcal{L}^2(\Omega;\mathbb{R})$

For any $1 \leq p < \infty$ let $\mathcal{L}^p(\Omega;\mathbb{R})$ [resp. $\mathcal{L}^p(\Omega_{\mathcal{F}};\mathbb{R})$] be the semi-normed linear space of all the real random variables on Ω [resp. on $\Omega_{\mathcal{F}}$] having finite moment of order p.

Corollary 880 *For any $X \in \mathcal{L}^p(\Omega;\mathbb{R})$, we have*

$$|\mathbf{E}[X \mid \mathcal{F}]|^p \leq \mathbf{E}[|X|^p \mid \mathcal{F}]. \tag{10.21}$$

Proof. The absolute value function $|\cdot| : \mathbb{R} \to \mathbb{R}$ is convex and $|X| \in \mathcal{L}^1(\Omega;\mathbb{R})$. Hence, applying (10.18), we obtain

$$|\mathbf{E}[X \mid \mathcal{F}]| \leq \mathbf{E}[|X| \mid \mathcal{F}]. \tag{10.22}$$

In addition, since for every $1 < p < \infty$ the function $\phi_p : \mathbb{R}_+ \to \mathbb{R}$ given by

$$\phi_p(x) \overset{\text{def}}{=} x^p, \quad \forall x \in \mathbb{R}$$

is increasing and convex and $|X|^p \in \mathcal{L}^1(\Omega;\mathbb{R})$, applying again (10.18) to $|X|$, we obtain

$$\mathbf{E}[|X| \mid \mathcal{F}]^p \leq \mathbf{E}[|X|^p \mid \mathcal{F}]. \tag{10.23}$$

In the end, combining (10.22) with (10.23), the desired (10.21) immediately follows. □

Corollary 881 *For any $X \in \mathcal{L}^p(\Omega;\mathbb{R})$ we have*

$$\mathbf{E}[|\mathbf{E}[X \mid \mathcal{F}]|^p] \leq \mathbf{E}[|X|^p]. \tag{10.24}$$

In particular, $\mathbf{E}[X \mid \mathcal{F}] \in \mathcal{L}^p(\Omega_{\mathcal{F}};\mathbb{R})$.

Proof. Applying the expectation operator to (10.21), on account of (10.13), we obtain

$$\mathbf{E}[|\mathbf{E}[X \mid \mathcal{F}]|^p] \leq \mathbf{E}[|X|^p], \tag{10.25}$$

which clearly implies (10.24). □

Corollary 882 *For any $X \in \mathcal{L}^2(\Omega;\mathbb{R})$ we have*

$$\mathbf{D}^2[\mathbf{E}[X \mid \mathcal{F}]] \leq \mathbf{D}^2[X].$$

Proof. As a particular case of (10.24), we have

$$\mathbf{E}\left[\mathbf{E}[X \mid \mathcal{F}]^2\right] \leq \mathbf{E}\left[X^2\right].$$

Hence, on account of (10.13), we obtain

$$\mathbf{D}^2[\mathbf{E}[X \mid \mathcal{F}]] = \mathbf{E}\left[\mathbf{E}[X \mid \mathcal{F}]^2\right] - \mathbf{E}[\mathbf{E}[X \mid \mathcal{F}]]^2 \leq \mathbf{E}\left[X^2\right] - \mathbf{E}[X]^2 = \mathbf{D}^2[X],$$

as desired. □

Theorem 883 (transparence property) *For all $X, Y \in \mathcal{L}^1(\Omega;\mathbb{R})$ such that $XY \in \mathcal{L}^1(\Omega;\mathbb{R})$ and $Y \in \mathcal{L}^1(\Omega_{\mathcal{F}};\mathbb{R})$ we have*

$$\mathbf{E}[XY \mid \mathcal{F}] = Y\mathbf{E}[X \mid \mathcal{F}].$$

In particular, Equation (10.17) holds true for every $X \in \mathcal{L}^p(\Omega;\mathbb{R})$ and every $Y \in \mathcal{L}^q(\Omega_{\mathcal{F}};\mathbb{R})$ such that $p^{-1} + q^{-1} = 1$.

Let $\mathcal{F},\mathcal{G}$ sub-σ-algebras of $\mathcal{E}$ such that $\mathcal{G}\subseteq\mathcal{F}$.

Corollary 884 (Further properties of the conditional expectation) *We have*

$$\mathbf{E}\left[XY^{\intercal}\mid\mathcal{F}\right]=\mathbf{E}\left[X\right]\mathbf{E}\left[Y^{\intercal}\mid\mathcal{F}\right]\qquad\textit{[resp. }\mathbf{E}\left[XY^{\intercal}\mid\mathcal{F}\right]=\mathbf{E}\left[X\mid\mathcal{F}\right]\mathbf{E}\left[Y\right]^{\intercal}\textit{]},\tag{10.26}$$

for every $X\in L^{2}\left(\Omega;\mathbb{R}^{M}\right)$ independent of $\mathcal{F}\vee\sigma(Y)$ and every $Y\in L^{2}\left(\Omega;\mathbb{R}^{N}\right)$ [resp. for every $X\in L^{2}\left(\Omega;\mathbb{R}^{M}\right)$ and every $Y\in L^{2}\left(\Omega;\mathbb{R}^{N}\right)$ independent of $\mathcal{F}\vee\sigma(X)$], where $\mathcal{F}\vee\sigma(Y)$ [resp. $\mathcal{F}\vee\sigma(X)$] is the σ-algebra generated by $\mathcal{F}$ and $\sigma(Y)$ [resp. $\sigma(X)$].

Proof. By virtue of the tower property (see Equation (10.15)), we can write

$$\mathbf{E}\left[XY^{\intercal}\mid\mathcal{F}\right]=\mathbf{E}\left[\mathbf{E}\left[XY^{\intercal}\mid\mathcal{F}\right]\mid\mathcal{F}\vee\sigma(Y)\right]=\mathbf{E}\left[\mathbf{E}\left[XY^{\intercal}\mid\mathcal{F}\vee\sigma(Y)\right]\mid\mathcal{F}\right].\tag{10.27}$$

On the other hand, since Y is clearly $(\mathcal{F}\vee\sigma(Y))$-measurable, by virtue of the transparence property (see Equation (10.17)) we have

$$\mathbf{E}\left[\mathbf{E}\left[XY^{\intercal}\mid\mathcal{F}\vee\sigma(Y)\right]\mid\mathcal{F}\right]=\mathbf{E}\left[\mathbf{E}\left[X\mid\mathcal{F}\vee\sigma(Y)\right]Y^{\intercal}\mid\mathcal{F}\right].\tag{10.28}$$

Now, since X is independent of $\mathcal{F}\vee\sigma(Y)$, by virtue of the concentration and linearity property (see Equations () and ()), we have

$$\mathbf{E}\left[\mathbf{E}\left[X\mid\mathcal{F}\vee\sigma(Y)\right]Y^{\intercal}\mid\mathcal{F}\right]=\mathbf{E}\left[\mathbf{E}\left[X\right]Y^{\intercal}\mid\mathcal{F}\right]=\mathbf{E}\left[X\right]\mathbf{E}\left[Y^{\intercal}\mid\mathcal{F}\right].\tag{10.29}$$

Cpmbining Equations (10.27)-(10.29) we obtain the first part of Equation (10.26). The proof of the second part is perfectly analogous. □

Let $1\leq p<\infty$ and let $L^{p}\left(\Omega;\mathbb{R}\right)$ [resp. $L^{p}(\Omega_{\mathcal{F}};\mathbb{R})$] be the Banach space of all the equivalence classes of real random variables on Ω [resp. on $\Omega_{\mathcal{F}}$] having finite moment of oreder p, modulus the $\mathbf{P}$-a.s. [resp. $\mathbf{P}_{|\mathcal{F}}$-a.s.] equality. We can think of $L^{p}(\Omega_{\mathcal{F}};\mathbb{R})$ as a subspace of $L^{p}\left(\Omega;\mathbb{R}\right)$. Moreover thanks to Corollary 881 we can consider the conditional expectation operator $\mathbf{E}\left[\cdot\mid\mathcal{F}\right]:L^{p}\left(\Omega;\mathbb{R}\right)\rightarrow L^{p}(\Omega_{\mathcal{F}};\mathbb{R})$. We have

Theorem 885 *For every $1\leq p\leq\infty$ the conditional expectation operator $\mathbf{E}\left[\cdot\mid\mathcal{F}\right]:L^{p}\left(\Omega;\mathbb{R}\right)\rightarrow L^{p}(\Omega_{\mathcal{F}};\mathbb{R})$ is a positive linear operator of unit norm. In particular, when $p=2$ it is an orthogonal projection and we have*

$$\mathbf{E}\left[X\mid\mathcal{F}\right]=\underset{Y\in L^{2}(\Omega_{\mathcal{F}};\mathbb{R})}{\arg\min}\ \mathbf{E}\left[\left(X-Y\right)^{2}\right].$$

Proof. see Chung (1974) [**?**, p. 301] □

10.2.4 Conditional Expectation Given a Random Variable

Let $(\Omega,\mathcal{E},\mathbf{P})\equiv\Omega$ be a probability space, let $(\mathbb{Y},\mathcal{N})\equiv\mathbb{Y}$ be a state space, let $Y:\Omega\rightarrow\mathbb{Y}$ be an $(\mathcal{E},\mathcal{N})$-random variable on Ω with states in $\mathbb{Y}$, and let $\sigma\left(Y\right)$ be the σ-algebra generated by Y.

Lemma 886 (Doob-Dynkin representation lemma) *Assume a function $Z:\Omega\rightarrow\mathbb{R}$ is a real $\sigma\left(Y\right)$-random variable. Then, there exists a function $\psi:\mathbb{Y}\rightarrow\mathbb{R}$ which is $(\mathcal{N},\mathcal{B}\left(\mathbb{R}\right))$-measurable and such that*

$$Z=\psi\left(Y\right).$$

Proof. Assume Z is simple, that is $Z(\Omega) \equiv (z_k)_{k=1}^n$, for some $n \in \mathbb{N}$ and $z_1, \ldots, z_n \in \mathbb{R}$. Then, setting $F_k \equiv \{Z = z_k\}$, for $k = 1, \ldots, n$, we can write

$$Z = \sum_{k=1}^{n} z_k 1_{F_k},$$

where $(F_k)_{k=1}^n$ is a partition of Ω of events in $\sigma(Y)$. We Then, have

$$F_k = \{Y \in N_k\},$$

where $N_{k=1}$ is a suitable subset of $\mathbb{Y}$ in $\mathcal{N}$, for every $k = 1, \ldots, n$. Note that the subsets in sequence $(N_k)_{k=1}^n$ do not need to be disjoint, unless Y is surjective. However, setting

$$M_1 \equiv N_1, \quad M_k \equiv N_k - \bigcup_{j=1}^{k-1} N_j, \quad \forall k = 1, \ldots, n,$$

we have that $(M_k)_{k=1}^n$ is a sequence of disjoint subsets of $\mathbb{Y}$ in $\mathcal{N}$ such that

$$\{Y \in M_k\} = \left\{ Y \in N_k - \bigcup_{j=1}^{k-1} N_j \right\} = \{Y \in N_k\} - \bigcup_{j=1}^{k-1} \{Y \in N_j\} = F_k - \bigcup_{j=1}^{k-1} \{Y \in F_j\} = F_k$$

and

$$1_{M_k}(Y(\omega)) = 1_{F_k}(\omega),$$

for every $k = 1, \ldots, n$, and every $\omega \in \Omega$. As a consequence, introducing the function $\psi : \mathbb{Y} \to \mathbb{R}$ given by

$$\psi(y) \overset{\text{def}}{=} \sum_{k=1}^{n} z_k 1_{M_k}(y), \quad \forall y \in \mathbb{Y},$$

we have that ψ is a simple $(\mathcal{N}, \mathcal{B}(\mathbb{R}))$-measurable function such that

$$\psi(Y(\omega)) = \sum_{k=1}^{n} z_k 1_{M_k}(Y(\omega)) = \sum_{k=1}^{n} z_k 1_{F_k}(\omega) = Z(\omega).$$

Therefore, the lemma is true in case Z is a simple real $\sigma(Y)$-random variable. On the other hand, we know that for every real $\sigma(Y)$-random variable Z is possible to determine a sequence $(Z_n)_{n \geq 1}$ of simple real $\sigma(Y)$-random variables such that

$$\lim_{n \to \infty} Z_n(\omega) = Z(\omega),$$

for every $\omega \in \Omega$. For what shown above, there exists a function $\psi_n : \mathbb{Y} \to \mathbb{R}$ which is $(\mathcal{N}, \mathcal{B}(\mathbb{R}))$-measurable and such that

$$Z_n = \psi_n(Y),$$

for every $n \geq 1$. Write

$$\mathbb{Y}_c \equiv \left\{ y \in \mathbb{Y} : \exists \lim_{n \to \infty} \psi_n(y) \right\},$$

it is well known that $\mathbb{Y}_c \in \mathcal{N}$. Furthermore, $Y(\omega) \in \mathbb{Y}_c$, for every $\omega \in \Omega$. Hence, introduce the function $\psi : \mathbb{Y} \to \mathbb{R}$ given by

$$\psi(y) \overset{\text{def}}{=} \left(\lim_{n \to \infty} \psi_n(y) \right) 1_{\mathbb{Y}_c}(y), \quad \forall y \in \mathbb{Y},$$

we have that ψ is a $(\mathcal{N}, \mathcal{B}(\mathbb{R}))$-measurable function such that

$$\psi(Y(\omega)) = \left(\lim_{n\to\infty} \psi_n(Y(\omega))\right) 1_{\mathbb{Y}_c}(Y(\omega)) = \lim_{n\to\infty} \psi_n(Y(\omega)) = \lim_{n\to\infty} Z_n(\omega) = Z(\omega).$$

This completes the proof (see Rao-Swift (2005) [6, Chapter 1, Proposition 3, p. 8] see also Lessi (1993) [5, Capitolo 3, Proposizione 3.3.21, p. 138]). □

As an applicazion of the Doob-Dynkin representation lemma to conditional expectation we have

Remark 887 *For every $X \in \mathcal{L}^1(\Omega;\mathbb{R})$ there exists a function $\psi : \mathbb{Y} \to \mathbb{R}$ which is $(\mathcal{N}, \mathcal{B}(\mathbb{R}))$-measurable and such that*

$$\mathbf{E}[X \mid Y] = \psi(Y).$$

In particular, if Y is discrete, that is to say $Y(\Omega) = \{y_n\}_{n\in N}$ for some $N \subseteq \mathbb{N}$, Then, we have

$$\mathbf{E}[X \mid Y] = \sum_{n\in N} \mathbf{E}[X \mid Y = y_n] 1_{\{Y=y_n\}},$$

which implies

$$\psi(y) = \sum_{n\in N} \mathbf{E}[X \mid Y = y_n] 1_{\{y_n\}}(y),$$

for every $y \in \mathbb{Y}$.

We have also,

Proposition 888 (Radon-Nikodým) *Let $\mathbf{P}_{|\sigma(Y)}$ be the restriction of the probability $\mathbf{P}$ to $\sigma(Y)$ and let $P_Y : \mathcal{N} \to \mathbb{R}$ be the distribution of Y. Then, the map $P_Y^X : \mathcal{N} \to \mathbb{R}$ given by*

$$P_Y^X(N) \overset{def}{=} \int_{\{Y\in N\}} X d\mathbf{P}_{|\sigma(Y)}, \quad \forall N \in \mathcal{N},$$

is a real measure on $\mathcal{N}$, which is absolutely continuous with respect to P_Y. Hence, there exists the Radon-Nikodým derivative $dP_Y^X/dP_Y : \mathbb{Y} \to \mathbb{R}$ such that

$$P_Y^X(N) = \int_N dP_Y^X/dP_Y(y)\, dP_Y(y),$$

for every $N \in \mathcal{N}$.

Definition 889 *The Radon-Nikodým derivative dP_Y^X/dP_Y is called the* conditional expectation *of X given $Y = y$ and is denoted by $\mathbf{E}[X \mid Y = y]$.*

Remark 890 *The conditional expectation $\mathbf{E}[X \mid Y = y] : \mathbb{Y} \to \mathbb{R}$ is an $(\mathcal{N}, \mathcal{B}(\mathbb{R}))$-measurable function such that*

$$\mathbf{E}[X \mid Y](\omega) = \mathbf{E}[X \mid Y = Y(\omega)],$$

$\mathbf{P}_{|\sigma(Y)}$-a.s. on $\Omega_{\sigma(Y)}$.

Proof. In fact, we have

$$\int_{\{Y\in N\}} \mathbf{E}\left[X \mid Y\right](\omega)\, d\mathbf{P}_{|\sigma(Y)}(\omega) = \int_{\{Y\in N\}} X d\mathbf{P}_{|\sigma(Y)} = P_Y^X(N)$$
$$= \int_N dP_Y^X/dP_Y(y)\, dP_Y(y) = \int_N \mathbf{E}\left[X \mid Y=y\right] dP_Y(y) = \int_{\{Y\in N\}} \mathbf{E}\left[X \mid Y=Y(\omega)\right] d\mathbf{P}_{|\sigma(Y)}(\omega),$$

for every $N \in \mathcal{N}$, which yields the desired result. □

Note that, combining Remarks (887) and (890) it clearly follows

$$\psi(y) = \mathbf{E}\left[X \mid Y=y\right],$$

P_Y-a.s. on $\mathbb{Y}$.

Proposition 891 *Assume $Y \in \mathcal{L}^1(\Omega;\mathbb{R})$ is absolutely continuous with strictly positive density. In addition, assume that X and Y are jointly absolutely continuous. Then, the conditional expectation $\mathbf{E}\left[X \mid Y=y\right]:\mathbb{R}\to\mathbb{R}$ is absolutely continuous and setting*

$$f_{X|Y}(x,y) \overset{def}{=} \frac{f_{X,Y}(x,y)}{f_Y(y)}, \quad \forall (x,y)\in\mathbb{R}^2, \tag{10.30}$$

where f_Y is the density of Y and $f_{X,Y}$ is the joint density of the random variables X and Y, we have

$$\mathbf{E}\left[X \mid Y=y\right] = \int_{\mathbb{R}} x f_{X|Y}(x,y)\, d\mu_L(x). \tag{10.31}$$

Proof. In fact, we have

$$\int_{\{Y\in B\}} \mathbf{E}\left[X \mid Y\right] d\mathbf{P}_{|\sigma(Y)} = \int_{\{Y\in B\}} X\, d\mathbf{P} = \int_\Omega 1_{\{Y\in B\}} X\, d\mathbf{P}$$
$$= \int_{\mathbb{R}}\int_{\mathbb{R}} 1_B(y)\, x f_{X,Y}(x,y)\, d\mu_{L^2}(x,y) = \int_{\mathbb{R}}\int_{\mathbb{R}} 1_B(y)\, x f_{X|Y}(x,y)\, f_Y(y)\, d\mu_{L^2}(x,y)$$
$$= \int_B \left(\int_{\mathbb{R}} x f_{X|Y}(x,y)\; d\mu_L(x)\right) f_Y(y)\, d\mu_L(y) = \int_B \left(\int_{\mathbb{R}} x f_{X|Y}(x,y)\; d\mu_L(x)\right) dP_Y$$

for every $B \in \mathcal{B}(\mathbb{R})$. On the other hand,

$$\int_{\{Y\in B\}} \mathbf{E}\left[X \mid Y\right] d\mathbf{P}_{|\sigma(Y)} = \int_B \mathbf{E}\left[X \mid Y=y\right] dP_Y(y).$$

as desired. □

Definition 892 *We call the* conditional density *of X given $Y=y$, the function $f_{X|Y}:\mathbb{R}^2\to\mathbb{R}_+$ introduced in Proposition 891.*

Proposition 893 *For any Borel function $h:\mathbb{R}\to\mathbb{R}$ such that $h\circ X \in \mathcal{L}^1(\Omega;\mathbb{R})$, we have*

$$\mathbf{E}\left[h(X) \mid Y=y\right] = \int_{\mathbb{R}} h(x) f_{X|Y}(x,y)\, d\mu_L(x). \tag{10.32}$$

Proof. . □

Proposition 894 *Assume X and Y are jointly normal real random variables with density*

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{X,Y}^2}} \exp\left(-\frac{1}{2(1-\rho_{X,Y}^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho_{X,Y}\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right), \tag{10.33}$$

where $\mu_X \equiv \mathbf{E}[X]$, $\mu_Y \equiv \mathbf{E}[Y]$, $\sigma_X \equiv \mathbf{D}[X]$, $\sigma_Y \equiv \mathbf{D}[Y]$, and $\rho_{X,Y} \equiv Corr(X,Y)$. Then, we have

$$\mathbf{E}[X \mid Y = y] = \mu_X + \rho_{X,Y}\frac{\sigma_X}{\sigma_Y}(y-\mu_Y) \tag{10.34}$$

and

$$\mathbf{E}\left[X^2 \mid Y = y\right] = \sigma_X^2(1-\rho_{X,Y}^2) + (\mu_X + \rho_{X,Y}\frac{\sigma_X}{\sigma_Y}(y-\mu_Y))^2. \tag{10.35}$$

As a consequence,

$$\mathbf{E}\left[X^2 \mid Y = y\right] - \mathbf{E}[X \mid Y = y]^2 = \sigma_X^2(1-\rho_{X,Y}^2). \tag{10.36}$$

Proof. Setting $\xi \equiv \frac{1}{\sqrt{2}\sigma_X\sqrt{1-\rho_{X,Y}^2}}\left(x - \mu_X - \rho_{X,Y}\frac{\sigma_X}{\sigma_Y}(y-\mu_Y)\right)$ a straightforward computation gives

$$\begin{aligned}
&f_Y(y)\\
&= \int_{-\infty}^{+\infty} f_{X,Y}(x,y)\,dx\\
&= \int_{-\infty}^{+\infty} \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{X,Y}^2}} \exp\left(-\frac{1}{2(1-\rho_{X,Y}^2)}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho_{X,Y}\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)\right.\right.\right.\\
&= \int_{-\infty}^{+\infty} \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{X,Y}^2}} \exp\left(-\frac{1}{2\sigma_X^2(1-\rho_{X,Y}^2)}\left((x-\mu_X)^2 - 2\rho_{X,Y}\frac{\sigma_X}{\sigma_Y}(x-\mu_X)(y-\mu_Y) + \frac{\sigma_X^2}{\sigma_Y^2}(y-\mu_\right.\right.\right.\\
&= \int_{-\infty}^{+\infty} \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{X,Y}^2}}\\
&\cdot \exp\left(-\frac{1}{2\sigma_X^2(1-\rho_{X,Y}^2)}\left((x-\mu_X)^2 - 2\rho_{X,Y}\frac{\sigma_X}{\sigma_Y}(x-\mu_X)(y-\mu_Y) + \rho_{X,Y}^2\frac{\sigma_X^2}{\sigma_Y^2}(y-\mu_Y)^2 + \frac{\sigma_X^2}{\sigma_Y^2}\left(1-\rho_{X,Y}^2\right)(y\right.\right.\right.\\
&= \exp\left(-\frac{1}{2\sigma_Y^2}(y-\mu_Y)^2\right)\int_{-\infty}^{+\infty} \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{X,Y}^2}} \exp\left(-\frac{1}{2\sigma_X^2(1-\rho_{X,Y}^2)}\left(x-\mu_X-\rho_{X,Y}\frac{\sigma_X}{\sigma_Y}y-\mu_Y\right)^2\right)\\
&= \frac{1}{\sqrt{2\pi}\sigma_Y}\exp\left(-\frac{1}{2\sigma_Y^2}(y-\mu_Y)^2\right)\int_{-\infty}^{+\infty}\exp\left(-\xi^2\right)\,d\xi\\
&= \frac{1}{\sqrt{2\pi}\sigma_Y}\exp\left(-\frac{1}{2\sigma_Y^2}(y-\mu_Y)^2\right).
\end{aligned}$$

As a consequence,

$$\frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{X,Y}^2}}\exp\left(-\frac{1}{2(1-\rho_{X,Y}^2)}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho_{X,Y}\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right)\right)}{\frac{1}{\sqrt{2\pi}\sigma_Y}\exp\left(-\frac{1}{2\sigma_Y^2}(y-\mu_Y)^2\right)}$$

$$= \frac{1}{\sqrt{2\pi}\sigma_X\sqrt{1-\rho_{X,Y}^2}}\exp\left(-\frac{1}{2\sigma_X^2(1-\rho_{X,Y}^2)}\left(x-\mu_X-\rho_{X,Y}\frac{\sigma_X}{\sigma_Y}(y-\mu_Y)\right)^2\right).$$

Still setting $\xi \equiv \frac{1}{\sqrt{2}\sigma_X\sqrt{1-\rho_{X,Y}^2}}\left(x-\mu_X-\rho_{X,Y}\frac{\sigma_X}{\sigma_Y}(y-\mu_Y)\right)$, the latter implies,

$$\begin{aligned}
\mathbf{E}[X \mid Y=y] &= \int_{-\infty}^{+\infty} x f_{X|Y}(x,y)\,dx \\
&= \int_{-\infty}^{+\infty} x\frac{f_{X,Y}(x,y)}{f_Y(y)}\,dx \\
&= \int_{-\infty}^{+\infty} \frac{x}{\sqrt{2\pi}\sigma_X\sqrt{1-\rho_{X,Y}^2}}\exp\left(-\frac{1}{2\sigma_X^2(1-\rho_{X,Y}^2)}\left(x-\mu_X-\rho_{X,Y}\frac{\sigma_X}{\sigma_Y}(y-\mu_Y)\right)^2\right)dx \\
&= \frac{1}{\sqrt{\pi}}\int_{-\infty}^{+\infty}\left(\sqrt{2}\sigma_X\sqrt{1-\rho_{X,Y}^2}\xi+\mu_X+\rho_{X,Y}\frac{\sigma_X}{\sigma_Y}(y-\mu_Y)\right)\exp\left(-\xi^2\right)d\xi \\
&= \frac{\sqrt{2}\sigma_X\sqrt{1-\rho_{X,Y}^2}}{\sqrt{\pi}}\int_{-\infty}^{+\infty}\xi\exp\left(-\xi^2\right)d\xi + \frac{\mu_X+\rho_{X,Y}\frac{\sigma_X}{\sigma_Y}(y-\mu_Y)}{\sqrt{\pi}}\int_{-\infty}^{+\infty}\exp\left(-\xi^2\right)d\xi \\
&= \mu_X+\rho_{X,Y}\frac{\sigma_X}{\sigma_Y}(y-\mu_Y),
\end{aligned}$$

which is the desired (10.34). Similarly, we have

$$\begin{aligned}
\mathbf{E}\left[X^2 \mid Y=y\right] &= \int_{-\infty}^{+\infty} x^2 f_{X|Y}(x,y)\,dx \\
&= \int_{-\infty}^{+\infty} x^2\frac{f_{X,Y}(x,y)}{f_Y(y)}\,dx \\
&= \int_{-\infty}^{+\infty} \frac{x^2}{\sqrt{2\pi}\sigma_X\sqrt{1-\rho_{X,Y}^2}}\exp\left(-\frac{1}{2\sigma_X^2(1-\rho_{X,Y}^2)}\left(x-\mu_X-\rho_{X,Y}\frac{\sigma_X}{\sigma_Y}(y-\mu_Y)\right)^2\right)dx \\
&= \frac{1}{\sqrt{\pi}}\int_{-\infty}^{+\infty}\left(\sqrt{2}\sigma_X\sqrt{1-\rho_{X,Y}^2}\xi+\mu_X+\rho_{X,Y}\frac{\sigma_X}{\sigma_Y}(y-\mu_Y)\right)^2\exp\left(-\xi^2\right)d\xi \\
&= \frac{2\sigma_X^2(1-\rho_{X,Y}^2)}{\sqrt{\pi}}\int_{-\infty}^{+\infty}\xi^2\exp\left(-\xi^2\right)d\xi \\
&+ \frac{2\sqrt{2}\sigma_X\sqrt{1-\rho_{X,Y}^2}(\mu_X+\rho_{X,Y}\frac{\sigma_X}{\sigma_Y}(y-\mu_Y))}{\sqrt{\pi}}\int_{-\infty}^{+\infty}\xi\exp\left(-\xi^2\right)d\xi \\
&+ \frac{(\mu_X+\rho_{X,Y}\frac{\sigma_X}{\sigma_Y}(y-\mu_Y))^2}{\sqrt{\pi}}\int_{-\infty}^{+\infty}\exp\left(-\xi^2\right)d\xi \\
&= \sigma_X^2(1-\rho_{X,Y}^2)+(\mu_X+\rho_{X,Y}\frac{\sigma_X}{\sigma_Y}(y-\mu_Y))^2,
\end{aligned}$$

which yields (10.35). □

Corollary 895 *Let X and Y be real random variables with joint normal distribution. Then, we have*

$$\begin{aligned}\mathbf{E}[X \mid Y] &= \mathbf{E}[X] + Corr(X,Y)\frac{\mathbf{D}[X]}{\mathbf{D}[Y]}(Y - \mathbf{E}[Y]) \\ &= \mathbf{E}[X] + \frac{Cov(X,Y)}{\mathbf{D}^2[Y]}(Y - \mathbf{E}[Y]) \end{aligned} \tag{10.37}$$

and

$$\begin{aligned}\mathbf{E}\left[X^2 \mid Y\right] &= \mathbf{D}^2[X]\left(1 - Corr(X,Y)^2\right) + \left(\mathbf{E}[X] + Corr(X,Y)\frac{\mathbf{D}[X]}{\mathbf{D}[Y]}(Y - \mathbf{E}[Y])\right)^2 \\ &= \mathbf{D}^2[X] - \frac{Cov(X,Y)^2}{\mathbf{D}^2[Y]} + \left(\mathbf{E}[X] + \frac{Cov(X,Y)}{\mathbf{D}^2[Y]}(Y - \mathbf{E}[Y])\right)^2. \end{aligned} \tag{10.38}$$

Let $X, Y \in \mathcal{L}^1(\Omega;\mathbb{R})$ such that $\mathbf{D}^2[Y] > 0$.

Theorem 896 *Assume $\mathbf{E}[X \mid Y]$ is linear in Y, that is $\mathbf{E}[X \mid Y] = a + bY$ for some $a, b \in \mathbb{R}$. Then, we have*

$$\begin{aligned}\mathbf{E}[X \mid Y] &= \mathbf{E}[X] + Corr(X,Y)\frac{\mathbf{D}[X]}{\mathbf{D}[Y]}(Y - \mathbf{E}[Y]) \\ &= \mathbf{E}[X] + \frac{Cov(X,Y)}{\mathbf{D}^2[Y]}(Y - \mathbf{E}[Y]). \end{aligned} \tag{10.39}$$

Proof. By computing the expectation of both sides of

$$\mathbf{E}[X \mid Y] = a + bX,$$

we obtain

$$\mathbf{E}[X] = \mathbf{E}[\mathbf{E}[X \mid Y]] = \mathbf{E}[a + bY] = a + b\mathbf{E}[Y]. \tag{10.40}$$

Moreover,

$$\mathbf{E}[XY] = \mathbf{E}[\mathbf{E}[XY \mid Y]] = \mathbf{E}[Y\mathbf{E}[X \mid Y]] = \mathbf{E}\left[aY + bY^2\right] = a\mathbf{E}[Y] + b\mathbf{E}\left[Y^2\right]. \tag{10.41}$$

Solving equations (10.40) and (10.41) in terms of a and b yields

$$a = \frac{\mathbf{E}[X]\mathbf{E}[Y^2] - \mathbf{E}[Y]\mathbf{E}[XY]}{\mathbf{D}^2[Y]},$$

and

$$b = \frac{\mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y]}{\mathbf{D}^2[Y]} = \frac{Cov(X,Y)}{\mathbf{D}^2[Y]}.$$

On the other hand, we can write

$$\begin{aligned}\frac{\mathbf{E}[X]\mathbf{E}[Y^2] - \mathbf{E}[Y]\mathbf{E}[XY]}{\mathbf{D}^2[Y]} &= \frac{\mathbf{E}[X]\mathbf{E}[Y^2] - \mathbf{E}[X]\mathbf{E}[Y]^2 + \mathbf{E}[X]\mathbf{E}[Y]^2 - \mathbf{E}[Y]\mathbf{E}[XY]}{\mathbf{D}^2[Y]} \\ &= \mathbf{E}[X] - \frac{\mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y]}{\mathbf{D}^2[Y]}\mathbf{E}[Y] \\ &= \mathbf{E}[X] - \frac{Cov(X,Y)}{\mathbf{D}^2[Y]}\mathbf{E}[Y], \end{aligned}$$

and the desired result immediately follows. □

Remark 897 *In terms of the conditional expectation of X given $Y = y$, to say the Borel measurable real function $\mathbf{E}[X \mid Y = y]$, we can write*

$$\mathbf{E}[X \mid Y = y] = \mathbf{E}[X] + \frac{Cov(X,Y)}{\mathbf{D}^2[Y]}(y - \mathbf{E}[Y]).$$

10.2.5 Conditional Independent Random Variables Given a σ-Field of Events

Let X, Y be real random variables on a probability space $\Omega \equiv (\Omega, \mathcal{E}, \mathbf{P})$ with finite moment of order one, and let $\mathcal{F}$ be a sub-sigma-algebra of $\mathcal{E}$.

Definition 898 *We say that X and Y are conditionally independent given $\mathcal{F}$, and write $X \perp\!\!\!\perp Y \mid \mathcal{F}$, if we have*

$$\mathbf{P}(U \cap V \mid \mathcal{F}) = \mathbf{P}(U \mid \mathcal{F})\,\mathbf{P}(V \mid \mathcal{F}), \quad \forall U \in \sigma(X), V \in \sigma(Y). \tag{10.42}$$

Equivalently,

$$\mathbf{P}(X \in A, Y \in B \mid \mathcal{F}) = \mathbf{P}(X \in A \mid \mathcal{F})\,\mathbf{P}(Y \in B \mid \mathcal{F}), \quad \forall A, B \in \mathcal{B}(\mathbb{R}). \tag{10.43}$$

Recall that

$$\mathbf{P}(E \mid \mathcal{F}) \equiv \mathbf{E}[1_E \mid \mathcal{F}], \quad \forall E \in \mathcal{E}.$$

Proposition 899 *Assume X and Y are conditionally independent given $\mathcal{F}$. Then, we have*

$$\mathbf{E}[XY^{\intercal} \mid \mathcal{F}] = \mathbf{E}[X \mid \mathcal{F}]\,\mathbf{E}[Y^{\intercal} \mid \mathcal{F}] \tag{10.44}$$

Proof. Assume $X \equiv 1_E$ and $Y \equiv 1_F$ where $E, F \in \mathcal{E}$. We have

$$\mathbf{E}[XY \mid \mathcal{F}] = \mathbf{E}[1_E 1_F \mid \mathcal{F}] = \mathbf{E}[1_{E\cap F} \mid \mathcal{F}] \equiv \mathbf{P}(E \cap F \mid \mathcal{F}). \tag{10.45}$$

On the other hand, we cleary have $E \in \sigma(X)$ and $F \in \sigma(Y)$. Hence, the assumption $X \perp\!\!\!\perp Y \mid \mathcal{F}$ implies

$$\mathbf{P}(E \cap F \mid \mathcal{F}) = \mathbf{P}(E \mid \mathcal{F})\,\mathbf{P}(F \mid \mathcal{F}) \equiv \mathbf{E}[1_E \mid \mathcal{F}]\,\mathbf{E}[1_F \mid \mathcal{F}] = \mathbf{E}[X \mid \mathcal{F}]\,\mathbf{E}[Y \mid \mathcal{F}]. \tag{10.46}$$

Combining (**??**) and (10.46), it Then, follows that (10.44) holds true for indicator functions of events... □

https://en.wikipedia.org/wiki/Conditional_independence

10.2.6 Conditional Variance Given a σ-Field of Events

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ a probability space, let $\mathcal{F}$ be a sub-sigma-algebra of $\mathcal{E}$, and let $L^2(\Omega; \mathbb{R}^N)$ [resp. $L^1(\Omega_{\mathcal{F}}; \mathbb{R}^N)$] the Hilbert space of all $\mathcal{E}$-measurable [resp. the Banach space of all $\mathcal{F}$-measurable] real N-dimensional random vectors, for some $N \in \mathbb{N}$, with finite moment of order 2.

Proposition 900 *Given any $X \in L^2(\Omega; \mathbb{R}^N)$, we call conditional variance of X given $\mathcal{F}$ the random vector $Var(X \mid \mathcal{F}) \in L^1(\Omega_{\mathcal{F}}; \mathbb{R}^N)$ given by*

$$Var(X \mid \mathcal{F}) \overset{def}{=} \mathbf{E}[(X - \mathbf{E}[X \mid \mathcal{F}])(X - \mathbf{E}[X \mid \mathcal{F}])^{\intercal} \mid \mathcal{F}]. \tag{10.47}$$

Definition 901 *In case $N = 1$, the conditional variance of X given $\mathcal{F}$ is also denoted by $\mathbf{D}^2[X \mid \mathcal{F}]$ and is clearly given by*

$$\mathbf{D}^2[X \mid \mathcal{F}] = \mathbf{E}\left[(X - \mathbf{E}[X \mid \mathcal{F}])^2 \mid \mathcal{F}\right]. \tag{10.48}$$

Remark 902 *We have*

$$Var(X \mid \mathcal{F}) = \mathbf{E}[XX^\intercal \mid \mathcal{F}] - \mathbf{E}[X \mid \mathcal{F}]\,\mathbf{E}[X^\intercal \mid \mathcal{F}]. \tag{10.49}$$

In particular,

$$\mathbf{D}^2[X \mid \mathcal{F}] = \mathbf{E}\left[X^2 \mid \mathcal{F}\right] - \mathbf{E}[X \mid \mathcal{F}]^2. \tag{10.50}$$

Proof. Thanks to Equations (10.14) and (10.13), a straightforward computation gives

$$\begin{aligned}
&\mathbf{E}[(X - \mathbf{E}[X \mid \mathcal{F}])(X - \mathbf{E}[X \mid \mathcal{F}])^\intercal \mid \mathcal{F}] \\
&= \mathbf{E}[(X - \mathbf{E}[X \mid \mathcal{F}])(X^\intercal - \mathbf{E}[X^\intercal \mid \mathcal{F}]) \mid \mathcal{F}] \\
&= \mathbf{E}[XX^\intercal - X\mathbf{E}[X^\intercal \mid \mathcal{F}] - \mathbf{E}[X \mid \mathcal{F}]X^\intercal + \mathbf{E}[X \mid \mathcal{F}]\,\mathbf{E}[X^\intercal \mid \mathcal{F}] \mid \mathcal{F}] \\
&= \mathbf{E}[XX^\intercal \mid \mathcal{F}] - \mathbf{E}[X\mathbf{E}[X^\intercal \mid \mathcal{F}] \mid \mathcal{F}] - \mathbf{E}[\mathbf{E}[X \mid \mathcal{F}]X^\intercal \mid \mathcal{F}] + \mathbf{E}[\mathbf{E}[X \mid \mathcal{F}]\,\mathbf{E}[X^\intercal \mid \mathcal{F}] \mid \mathcal{F}] \\
&= \mathbf{E}[XX^\intercal \mid \mathcal{F}] - \mathbf{E}[X \mid \mathcal{F}]\,\mathbf{E}[X^\intercal \mid \mathcal{F}] - \mathbf{E}[X \mid \mathcal{F}]\,\mathbf{E}[X^\intercal \mid \mathcal{F}] + \mathbf{E}[X \mid \mathcal{F}]\,\mathbf{E}[X^\intercal \mid \mathcal{F}] \\
&= \mathbf{E}[XX^\intercal \mid \mathcal{F}] - \mathbf{E}[X \mid \mathcal{F}]\,\mathbf{E}[X^\intercal \mid \mathcal{F}],
\end{aligned}$$

Equation (10.69) clearly follows. □

Remark 903 *Assume X is $\mathcal{F}$-measurable. Then*

$$Var(X \mid \mathcal{F}) = 0. \tag{10.51}$$

Remark 904 *Assume X is $\mathcal{F}$-independent. Then*

$$Var(X \mid \mathcal{F}) = Var(X). \tag{10.52}$$

Proposition 905 *Assume X is $\mathcal{F}$-measurable and Y is $\mathcal{F}$-independent. Then*

$$Var(X + Y \mid \mathcal{F}) = Var(Y). \tag{10.53}$$

Proof. Considering Equation (10.49), we can write

$$\begin{aligned}
&Var(X + Y \mid \mathcal{F}) \\
&= \mathbf{E}[(X + Y)(X + Y)^\intercal \mid \mathcal{F}] - \mathbf{E}[X + Y \mid \mathcal{F}]\,\mathbf{E}[(X + Y)^\intercal \mid \mathcal{F}] \\
&= \mathbf{E}[XX^\intercal + XY^\intercal + YX^\intercal + YY^\intercal \mid \mathcal{F}] - (\mathbf{E}[X \mid \mathcal{F}] + \mathbf{E}[Y \mid \mathcal{F}])(\mathbf{E}[X^\intercal \mid \mathcal{F}] + \mathbf{E}[Y^\intercal \mid \mathcal{F}]) \\
&= \mathbf{E}[XX^\intercal \mid \mathcal{F}] + \mathbf{E}[XY^\intercal \mid \mathcal{F}] + \mathbf{E}[YX^\intercal \mid \mathcal{F}] + \mathbf{E}[YY^\intercal \mid \mathcal{F}] \\
&- (\mathbf{E}[X \mid \mathcal{F}]\,\mathbf{E}[X^\intercal \mid \mathcal{F}] + \mathbf{E}[X \mid \mathcal{F}]\,\mathbf{E}[Y^\intercal \mid \mathcal{F}] + \mathbf{E}[Y \mid \mathcal{F}]\,\mathbf{E}[X^\intercal \mid \mathcal{F}] + \mathbf{E}[Y \mid \mathcal{F}]\,\mathbf{E}[Y^\intercal \mid \mathcal{F}]) \\
&= (\mathbf{E}[XX^\intercal \mid \mathcal{F}] - \mathbf{E}[X \mid \mathcal{F}]\,\mathbf{E}[X^\intercal \mid \mathcal{F}]) + (\mathbf{E}[XY^\intercal \mid \mathcal{F}] - \mathbf{E}[X \mid \mathcal{F}]\,\mathbf{E}[Y^\intercal \mid \mathcal{F}]) \\
&+ (\mathbf{E}[YX^\intercal \mid \mathcal{F}] - \mathbf{E}[Y \mid \mathcal{F}]\,\mathbf{E}[X^\intercal \mid \mathcal{F}]) + (\mathbf{E}[YY^\intercal \mid \mathcal{F}] - \mathbf{E}[Y \mid \mathcal{F}]\,\mathbf{E}[Y^\intercal \mid \mathcal{F}]) \\
&= Var(X \mid \mathcal{F}) + Var(Y \mid \mathcal{F}) \\
&+ (\mathbf{E}[XY^\intercal \mid \mathcal{F}] - \mathbf{E}[X \mid \mathcal{F}]\,\mathbf{E}[Y^\intercal \mid \mathcal{F}]) + (\mathbf{E}[YX^\intercal \mid \mathcal{F}] - \mathbf{E}[Y \mid \mathcal{F}]\,\mathbf{E}[X^\intercal \mid \mathcal{F}])
\end{aligned} \tag{10.54}$$

On the other hand, since X is $\mathcal{F}$-measurable and Y is $\mathcal{F}$-independent, we have

$$\begin{aligned}\mathbf{E}\left[XY^{\intercal} \mid \mathcal{F}\right] - \mathbf{E}\left[X \mid \mathcal{F}\right]\mathbf{E}\left[Y^{\intercal} \mid \mathcal{F}\right] &= X\mathbf{E}\left[Y^{\intercal} \mid \mathcal{F}\right] - \mathbf{E}\left[X \mid \mathcal{F}\right]\mathbf{E}\left[Y^{\intercal} \mid \mathcal{F}\right] \\ &= \left(X - \mathbf{E}\left[X \mid \mathcal{F}\right]\right)\mathbf{E}\left[Y^{\intercal}\right] = \\ &= \left(\mathbf{E}\left[X \mid \mathcal{F}\right] - \mathbf{E}\left[X \mid \mathcal{F}\right]\right)\mathbf{E}\left[Y^{\intercal}\right] \\ &= 0. \end{aligned} \tag{10.55}$$

Similarly

$$\begin{aligned}\mathbf{E}\left[YX^{\intercal} \mid \mathcal{F}\right] - \mathbf{E}\left[Y \mid \mathcal{F}\right]\mathbf{E}\left[X^{\intercal} \mid \mathcal{F}\right] &= \mathbf{E}\left[Y \mid \mathcal{F}\right]X^{\intercal} - \mathbf{E}\left[Y \mid \mathcal{F}\right]\mathbf{E}\left[X^{\intercal} \mid \mathcal{F}\right] \\ &= \mathbf{E}\left[Y\right]\left(X^{\intercal} - \mathbf{E}\left[X^{\intercal} \mid \mathcal{F}\right]\right) \\ &= \mathbf{E}\left[Y\right]\left(\mathbf{E}\left[X^{\intercal} \mid \mathcal{F}\right] - \mathbf{E}\left[X^{\intercal} \mid \mathcal{F}\right]\right) \\ &= 0. \end{aligned} \tag{10.56}$$

Combining Equations (10.54)-(10.56), on account of (10.51) and (10.53), the desired (10.53) follows.

Proposition 906 *Assume X is $\mathcal{F}$-measurable or X is $(\mathcal{F} \vee \sigma(Y))$-independent[2]. Then*

$$Var\left(X + Y \mid \mathcal{F}\right) = Var\left(X \mid \mathcal{F}\right) + Var\left(Y \mid \mathcal{F}\right). \tag{10.57}$$

Proof. Considering the proof of Proposition 905, Equation (10.57) will be proved if we show that (10.55) and (10.56) still holds true. Now, in case X is $\mathcal{F}$-measurable, we have

$$\begin{aligned}\mathbf{E}\left[XY^{\intercal} \mid \mathcal{F}\right] - \mathbf{E}\left[X \mid \mathcal{F}\right]\mathbf{E}\left[Y^{\intercal} \mid \mathcal{F}\right] &= X\mathbf{E}\left[Y \mid \mathcal{F}\right] - \mathbf{E}\left[X \mid \mathcal{F}\right]\mathbf{E}\left[Y \mid \mathcal{F}\right] \\ &= \mathbf{E}\left[X \mid \mathcal{F}\right]\mathbf{E}\left[Y \mid \mathcal{F}\right] - \mathbf{E}\left[X \mid \mathcal{F}\right]\mathbf{E}\left[Y \mid \mathcal{F}\right] \\ &= 0 \end{aligned}$$

and

$$\begin{aligned}\mathbf{E}\left[YX^{\intercal} \mid \mathcal{F}\right] - \mathbf{E}\left[Y \mid \mathcal{F}\right]\mathbf{E}\left[X^{\intercal} \mid \mathcal{F}\right] &= \mathbf{E}\left[Y \mid \mathcal{F}\right]X^{\intercal} - \mathbf{E}\left[Y \mid \mathcal{F}\right]\mathbf{E}\left[X^{\intercal} \mid \mathcal{F}\right] \\ &= \mathbf{E}\left[Y \mid \mathcal{F}\right]\mathbf{E}\left[X^{\intercal} \mid \mathcal{F}\right] - \mathbf{E}\left[Y \mid \mathcal{F}\right]\mathbf{E}\left[X^{\intercal} \mid \mathcal{F}\right] \\ &= 0. \end{aligned}$$

In case X is $(\mathcal{F} \vee \sigma(Y))$-independent, we have

$$\mathbf{E}\left[XY^{\intercal} \mid \mathcal{F}\right] - \mathbf{E}\left[X \mid \mathcal{F}\right]\mathbf{E}\left[Y^{\intercal} \mid \mathcal{F}\right] = \mathbf{E}\left[X\right]\mathbf{E}\left[Y^{\intercal} \mid \mathcal{F}\right] - \mathbf{E}\left[X\right]\mathbf{E}\left[Y^{\intercal} \mid \mathcal{F}\right] = 0$$

and

$$\mathbf{E}\left[YX^{\intercal} \mid \mathcal{F}\right] - \mathbf{E}\left[Y \mid \mathcal{F}\right]\mathbf{E}\left[X^{\intercal} \mid \mathcal{F}\right] = \mathbf{E}\left[Y \mid \mathcal{F}\right]\mathbf{E}\left[X^{\intercal}\right] - \mathbf{E}\left[Y \mid \mathcal{F}\right]\mathbf{E}\left[X^{\intercal}\right] = 0.$$

Therefore, the desired result follows. □

Proposition 907 *We have*

$$\mathbf{E}\left[Var\left(X \mid \mathcal{F}\right)\right] = \mathbf{E}\left[\left(X - \mathbf{E}\left[X \mid \mathcal{F}\right]\right)\left(X^{\intercal} - \mathbf{E}\left[X^{\intercal} \mid \mathcal{F}\right]\right)\right] \tag{10.58}$$

In particular,

$$\mathbf{E}\left[\mathbf{D}^2\left[X \mid \mathcal{F}\right]\right] = \mathbf{E}\left[\left(X - \mathbf{E}\left[X \mid \mathcal{F}\right]\right)^2\right]. \tag{10.59}$$

[2]In general, it is not enough to assume that X is independent of $\mathcal{F}$ and $\sigma(Y)$.

Proof. Thanks to Equation (10.13) we have

$$\mathbf{E}\left[Var\left(X \mid \mathcal{F}\right)\right] = \mathbf{E}\left[\mathbf{E}\left[\left(X - \mathbf{E}\left[X \mid \mathcal{F}\right]\right)\left(X - \mathbf{E}\left[X \mid \mathcal{F}\right]\right)^{\intercal} \mid \mathcal{F}\right]\right] = \mathbf{E}\left[\left(X - \mathbf{E}\left[X \mid \mathcal{F}\right]\right)\left(X^{\intercal} - \mathbf{E}\left[X^{\intercal} \mid \mathcal{F}\right]\right)\right],$$

as desired. □

Corollary 908 *We have*

$$\left\|X - \mathbf{E}\left[X \mid \mathcal{F}\right]\right\|_2 = \operatorname{tr}\left(\mathbf{E}\left[Var\left(X \mid \mathcal{F}\right)\right]\right). \tag{10.60}$$

In words: the trace of the expectation of the conditional variance of a random vector yields the mean-square distance between the random vector and its conditional expectation.

Proof. Since the trace functional commutes with expectation operator, considering Equation (10.58), we can write

$$\begin{aligned}\left\|X - \mathbf{E}\left[X \mid \mathcal{F}\right]\right\|_2 &= \mathbf{E}\left[\left(X - \mathbf{E}\left[X \mid \mathcal{F}\right]\right)^{\intercal}\left(X - \mathbf{E}\left[X \mid \mathcal{F}\right]\right)\right] \\ &= \mathbf{E}\left[\operatorname{tr}\left(\left(X - \mathbf{E}\left[X \mid \mathcal{F}\right]\right)\left(X - \mathbf{E}\left[X \mid \mathcal{F}\right]\right)^{\intercal}\right)\right] \\ &= \operatorname{tr}\mathbf{E}\left[\left(X - \mathbf{E}\left[X \mid \mathcal{F}\right]\right)\left(X - \mathbf{E}\left[X \mid \mathcal{F}\right]\right)^{\intercal}\right] \\ &= \operatorname{tr}\mathbf{E}\left[Var\left(X \mid \mathcal{F}\right)\right].\end{aligned}$$

This proves the claim. □

Proposition 909 (law of total variance) *We have*

$$Var\left(X\right) = \mathbf{E}\left[Var\left(X \mid \mathcal{F}\right)\right] + Var\left(\mathbf{E}\left[X \mid \mathcal{F}\right]\right) \tag{10.61}$$

In particular,

$$\mathbf{D}^2\left[X\right] = \mathbf{E}\left[\mathbf{D}^2\left[X \mid \mathcal{F}\right]\right] + \mathbf{D}^2\left[\mathbf{E}\left[X \mid \mathcal{F}\right]\right] \tag{10.62}$$

The result formulated by Equation (10.61) is termed as *law of total variance.*

Proof. On account of (10.49) and (10.13), a straightforward computation gives

$$\begin{aligned}\mathbf{E}\left[XX^{\intercal}\right] - \mathbf{E}\left[X\right]\mathbf{E}\left[X^{\intercal}\right] &= \mathbf{E}\left[\mathbf{E}\left[XX^{\intercal} \mid \mathcal{F}\right]\right] - \mathbf{E}\left[\mathbf{E}\left[X \mid \mathcal{F}\right]\right]\mathbf{E}\left[\mathbf{E}\left[X^{\intercal} \mid \mathcal{F}\right]\right] \\ &= \mathbf{E}\left[Var\left(X \mid \mathcal{F}\right) + \mathbf{E}\left[X \mid \mathcal{F}\right]\mathbf{E}\left[X^{\intercal} \mid \mathcal{F}\right]\right] - \mathbf{E}\left[\mathbf{E}\left[X \mid \mathcal{F}\right]\right]\mathbf{E}\left[\mathbf{E}\left[X^{\intercal} \mid \mathcal{F}\right]\right] \\ &= \mathbf{E}\left[Var\left(X \mid \mathcal{F}\right)\right] + \mathbf{E}\left[\mathbf{E}\left[X \mid \mathcal{F}\right]\mathbf{E}\left[X^{\intercal} \mid \mathcal{F}\right]\right] - \mathbf{E}\left[\mathbf{E}\left[X \mid \mathcal{F}\right]\right]\mathbf{E}\left[\mathbf{E}\left[X^{\intercal} \mid \mathcal{F}\right]\right] \\ &= \mathbf{E}\left[Var\left(X \mid \mathcal{F}\right)\right] + \mathbf{E}\left[\mathbf{E}\left[X \mid \mathcal{F}\right]\mathbf{E}\left[X \mid \mathcal{F}\right]^{\intercal}\right] - \mathbf{E}\left[\mathbf{E}\left[X \mid \mathcal{F}\right]\right]\mathbf{E}\left[\mathbf{E}\left[X \mid \mathcal{F}\right]^{\intercal}\right] \\ &= \mathbf{E}\left[Var\left(X \mid \mathcal{F}\right)\right] + Var\left(\mathbf{E}\left[X \mid \mathcal{F}\right]\right).\end{aligned}$$

which clearly yields the desired claim. □

Corollary 910 (law of total variance) *We have*

$$\left\|X - \mathbf{E}\left[X \mid \mathcal{F}\right]\right\|_2 = \operatorname{tr}\left(Var\left(X\right)\right) - \operatorname{tr}\left(Var\left(\mathbf{E}\left[X \mid \mathcal{F}\right]\right)\right). \tag{10.63}$$

In words: the larger is the trace of the variance of $\mathbf{E}\left[X \mid \mathcal{F}\right]$ *the better is* $\mathbf{E}\left[X \mid \mathcal{F}\right]$ *as the maximum mean square estimator of* X*.*

Proof. Combining Equations (10.60) and (10.61), we can write

$$\mathrm{tr}\left(Var\left(X\right)\right)=\mathrm{tr}\left(\mathbf{E}\left[Var\left(X\mid\mathcal{F}\right)\right]\right)+\mathrm{tr}\left(Var\left(\mathbf{E}\left[X\mid\mathcal{F}\right]\right)\right)=\left\|X-\mathbf{E}\left[X\mid\mathcal{F}\right]\right\|_{2}+\mathrm{tr}\left(Var\left(\mathbf{E}\left[X\mid\mathcal{F}\right]\right)\right)$$

and Equation (10.63) immediately follows. □

Proposition 911 *Assume X and Y are jointly normal real random variables. Then, we have*

$$\mathbf{D}^{2}\left[X\mid Y\right]=\mathbf{D}^{2}\left[X\right]\left(1-Corr(X,Y)^{2}\right)=\mathbf{D}^{2}\left[X\right]-\frac{Cov(X,Y)^{2}}{\mathbf{D}^{2}\left[Y\right]}. \tag{10.64}$$

Proof. By virtue of (10.50), considering Equations (**??**) and (**??**), we can write

$$\begin{aligned}\mathbf{D}^{2}\left[X\mid Y\right]&=\mathbf{E}\left[X^{2}\mid Y\right]-\mathbf{E}\left[X\mid Y\right]^{2}\\&=\mathbf{D}^{2}\left[X\right]\left(1-\frac{Cov(X,Y)^{2}}{\mathbf{D}^{2}\left[X\right]\mathbf{D}^{2}\left[Y\right]}\right)+\left(\mathbf{E}\left[X\right]+\frac{Cov(X,Y)}{\mathbf{D}^{2}\left[Y\right]}\left(Y-\mathbf{E}\left[Y\right]\right)\right)^{2}-\left(\mathbf{E}\left[X\right]+\frac{Cov(X,Y)}{\mathbf{D}^{2}\left[Y\right]}\left(\right.\right.\\&=\mathbf{D}^{2}\left[X\right]\left(1-\frac{Cov(X,Y)^{2}}{\mathbf{D}^{2}\left[X\right]\mathbf{D}^{2}\left[Y\right]}\right),\end{aligned}$$

which is the derired result. □

Equation (10.64) should be compared with (10.36) of Proposition 894.

10.2.7 Conditional Covariance Given a σ-Field of Events

Let $(\Omega,\mathcal{E},\mathbf{P})\equiv\Omega$ a probability space, let $\mathcal{F}$ be a sub-sigma-algebra of $\mathcal{E}$, and let $L^{2}\left(\Omega;\mathbb{R}^{M}\right)$ and $L^{2}\left(\Omega;\mathbb{R}^{N}\right)$ [resp. $L^{2}\left(\Omega_{\mathcal{F}};\mathbb{R}^{M}\right)$ and $L^{2}\left(\Omega_{\mathcal{F}};\mathbb{R}^{N}\right)$] the Hilbert space of all $\mathcal{E}$-measurable [resp. $\mathcal{F}$-measurable] real M-dimensional and N-dimensional random vectors, for some $M,N\in\mathbb{N}$, with finite moment of order 2.

Definition 912 *Given any $X\in L^{2}\left(\Omega;\mathbb{R}^{M}\right)$ and $Y\in L^{2}\left(\Omega;\mathbb{R}^{N}\right)$, we call the conditional covariance of X and Y given $\mathcal{F}$ the random matrix $Cov\left(X,Y\mid\mathcal{F}\right)\in L^{1}\left(\Omega_{\mathcal{F}};\mathbb{R}^{M\times N}\right)$*

$$Cov\left(X,Y\mid\mathcal{F}\right)\overset{def}{=}\mathbf{E}\left[\left(X-\mathbf{E}\left[X\mid\mathcal{F}\right]\right)\left(Y-\mathbf{E}\left[Y\mid\mathcal{F}\right]\right)^{\intercal}\mid\mathcal{F}\right].$$

Note that in case $Y=X$, we clearly have

$$Cov\left(X,Y\mid\mathcal{F}\right)=Var\left(X\mid\mathcal{F}\right).$$

Proposition 913 *We have*

$$Cov\left(X,Y\mid\mathcal{F}\right)=\mathbf{E}\left[XY^{\intercal}\mid\mathcal{F}\right]-\mathbf{E}\left[X\mid\mathcal{F}\right]\mathbf{E}\left[Y^{\intercal}\mid\mathcal{F}\right]. \tag{10.65}$$

Proof. Thanks to Equations (10.14) and (10.13), a straightforward computation gives

$$\begin{aligned}&\mathbf{E}\left[\left(X-\mathbf{E}\left[X\mid\mathcal{F}\right]\right)\left(Y-\mathbf{E}\left[Y\mid\mathcal{F}\right]\right)^{\intercal}\mid\mathcal{F}\right]\\&=\mathbf{E}\left[\left(X-\mathbf{E}\left[X\mid\mathcal{F}\right]\right)\left(Y^{\intercal}-\mathbf{E}\left[Y^{\intercal}\mid\mathcal{F}\right]\right)\mid\mathcal{F}\right]\\&=\mathbf{E}\left[XY^{\intercal}-X\mathbf{E}\left[Y^{\intercal}\mid\mathcal{F}\right]-\mathbf{E}\left[X\mid\mathcal{F}\right]Y^{\intercal}+\mathbf{E}\left[X\mid\mathcal{F}\right]\mathbf{E}\left[Y^{\intercal}\mid\mathcal{F}\right]\mid\mathcal{F}\right]\\&=\mathbf{E}\left[XY^{\intercal}\mid\mathcal{F}\right]-\mathbf{E}\left[X\mathbf{E}\left[Y^{\intercal}\mid\mathcal{F}\right]\mid\mathcal{F}\right]-\mathbf{E}\left[\mathbf{E}\left[X\mid\mathcal{F}\right]Y^{\intercal}\mid\mathcal{F}\right]+\mathbf{E}\left[\mathbf{E}\left[X\mid\mathcal{F}\right]\mathbf{E}\left[Y^{\intercal}\mid\mathcal{F}\right]\mid\mathcal{F}\right]\\&=\mathbf{E}\left[XY^{\intercal}\mid\mathcal{F}\right]-\mathbf{E}\left[X\mid\mathcal{F}\right]\mathbf{E}\left[Y^{\intercal}\mid\mathcal{F}\right]-\mathbf{E}\left[X\mid\mathcal{F}\right]\mathbf{E}\left[Y^{\intercal}\mid\mathcal{F}\right]+\mathbf{E}\left[X\mid\mathcal{F}\right]\mathbf{E}\left[Y^{\intercal}\mid\mathcal{F}\right]\\&=\mathbf{E}\left[XY^{\intercal}\mid\mathcal{F}\right]-\mathbf{E}\left[X\mid\mathcal{F}\right]\mathbf{E}\left[Y^{\intercal}\mid\mathcal{F}\right],\end{aligned}$$

as desired.

By wirtue of Equations (10.44) and (10.65), it immediately follows

Remark 914 *If the random vectors X and Y are conditionally independent of $\mathcal{F}$, then*

$$Cov\left(X, Y \mid \mathcal{F}\right) = Cov\left(Y, X \mid \mathcal{F}\right) = 0. \tag{10.66}$$

Proposition 915 *Assume X and Y are $\mathcal{F}$-conditionally independent. Then, we have*

$$Var\left(X + Y \mid \mathcal{F}\right) = Var\left(X \mid \mathcal{F}\right) + Var\left(Y \mid \mathcal{F}\right). \tag{10.67}$$

Proof. In light of Equation (10.65), we can rewrite Equation (10.54) in the form

$$Var\left(X + Y \mid \mathcal{F}\right) = Var\left(X \mid \mathcal{F}\right) + Var\left(Y \mid \mathcal{F}\right) + Cov\left(X, Y \mid \mathcal{F}\right) + Cov\left(Y, X \mid \mathcal{F}\right) \tag{10.68}$$

On the other hand, since X and Y are $\mathcal{F}$-conditionally independent, Equation (10.66) holds true. Combining Equations (10.68) and (10.66) the desired (10.67) immediately follows. □

Proposition 916 (Law of Total Covariance) *We have*

$$Cov\left(X, Y\right) = \mathbf{E}\left[Cov\left(X, Y \mid \mathcal{F}\right)\right] + Cov\left(\mathbf{E}\left[X \mid \mathcal{F}\right], \mathbf{E}\left[Y \mid \mathcal{F}\right]\right). \tag{10.69}$$

Proof. We have

$$Cov\left(X, Y\right) = \mathbf{E}\left[XY^{\intercal}\right] - \mathbf{E}\left[X\right]\mathbf{E}\left[Y^{\intercal}\right]$$

(see Equation ...). On the other hand, thanks to Equations (10.13) and (**??**), we can write

$$\begin{aligned}
\mathbf{E}\left[XY^{\intercal}\right] - \mathbf{E}\left[X\right]\mathbf{E}\left[Y^{\intercal}\right] &= \mathbf{E}\left[\mathbf{E}\left[XY^{\intercal} \mid \mathcal{F}\right]\right] - \mathbf{E}\left[\mathbf{E}\left[X \mid \mathcal{F}\right]\right]\mathbf{E}\left[\mathbf{E}\left[Y^{\intercal} \mid \mathcal{F}\right]\right] \\
&= \mathbf{E}\left[Cov\left(X, Y \mid \mathcal{F}\right) + \mathbf{E}\left[X \mid \mathcal{F}\right]\mathbf{E}\left[Y^{\intercal} \mid \mathcal{F}\right]\right] - \mathbf{E}\left[\mathbf{E}\left[X \mid \mathcal{F}\right]\right]\mathbf{E}\left[\mathbf{E}\left[Y^{\intercal} \mid \mathcal{F}\right]\right] \\
&= \mathbf{E}\left[Cov\left(X, Y \mid \mathcal{F}\right)\right] + \mathbf{E}\left[\mathbf{E}\left[X \mid \mathcal{F}\right]\mathbf{E}\left[Y^{\intercal} \mid \mathcal{F}\right]\right] - \mathbf{E}\left[\mathbf{E}\left[X \mid \mathcal{F}\right]\right]\mathbf{E}\left[\mathbf{E}\left[Y^{\intercal} \mid \mathcal{F}\right]\right] \\
&= \mathbf{E}\left[Cov\left(X, Y \mid \mathcal{F}\right)\right] + Cov\left(\mathbf{E}\left[X \mid \mathcal{F}\right], \mathbf{E}\left[Y \mid \mathcal{F}\right]\right).
\end{aligned}$$

Equation (10.69) clearly follows.

Chapter 11

Sequences of Real Random Variables

In this chapter we introduce the main converge properties of sequences of random variables.

11.1 Modes of Convergence

11.1.1 Almost Sure Convergence

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space and let $(X_n)_{n\geq 1}$ [resp. X] be a sequence of real random variables [resp. a real random variable] on Ω.

Definition 917 *We say that the sequence* $(X_n)_{n\geq 1}$ converges almost surely *to* X *(as* n *goes to infinity), and we write*

$$X_n \overset{a.s.}{\to} X,$$

if there exists an almost sure event $E \in \mathcal{E}$ *such that*

$$\lim_{n\to\infty} X_n(\omega) = X(\omega),$$

for every $\omega \in E$.

Example 918 *Consider the Borel-Lebesque probability space* $([0,1], \mathcal{B}([0,1]), \mu_L) \equiv \Omega$. *Write* $(q_n)_{n\geq 1}$ *for the sequence of all rational numbers in* $[0,1]$ *and consider the sequence* $(X_n)_{n\geq 1}$ *of real random variables on* Ω *given by*

$$X_n(\omega) \overset{def}{=} \begin{cases} (-1)^n, & \text{if } \omega = q_1, \ldots, q_n, \\ 1/n, & \text{if } \omega \in \Omega - \{q_1, \ldots, q_n\}. \end{cases}$$

We have $X_n \overset{a.s.}{\to} 0$, *where* 0 *stands for the Dirac random variable on* Ω *concentrated at* 0.

Discussion. Setting $E = [0,1] - \mathbb{Q}$, we have $\mathbf{P}(E) = 1$ and

$$\lim_{n\to\infty} X_n(\omega) = 0 = 0(\omega),$$

for every $\omega \in E$. □

Remark 919 *The sequence $(X_n)_{n\geq 1}$ converges almost surely to X, if and only if there exists an almost sure event $E \in \mathcal{E}$ such that, for any $\omega \in E$ and any $\varepsilon > 0$, we have*

$$|X_n(\omega) - X(\omega)| < \varepsilon,$$

for every $n \geq n(\omega, \varepsilon)$, where $n(\omega, \varepsilon)$ is a suitable positive integer depending on ω and ε.

Remark 920 *The sequence $(X_n)_{n\geq 1}$ converges almost surely to X,if and only if there exists an almost impossible event $F \in \mathcal{E}$ such that*

$$\lim_{n\to\infty} X_n(\omega) = X(\omega),$$

for every $\omega \in \Omega - F$.

Remark 921 *The sequence $(X_n)_{n\geq 1}$ converges almost surely to X, if and only if the sequence $(\tilde{X}_n)_{n\geq 1}$ given by*

$$\tilde{X}_n \overset{def}{=} X_n - X, \quad \forall n \geq 1.$$

converges almost surely to zero, to say $(\tilde{X}_n)_{n\geq 1}$ converges almost surely to the Dirac random variable concentrated at 0. In symbols

$$X_n \overset{\boldsymbol{a.s.}}{\to} X \Leftrightarrow X_n - X \overset{\boldsymbol{a.s.}}{\to} 0.$$

Theorem 922 (Cauchy criterion) *The sequence $(X_n)_{n\geq 1}$ converges almost surely, if and only if there exists an almost sure event E such that, for any $\omega \in \mathcal{E}$ and any $\varepsilon > 0$, we have*

$$|X_n(\omega) - X_m(\omega)| < \varepsilon,$$

for all $n, m \geq n(\omega, \varepsilon)$, where $n(\omega, \varepsilon)$ is a suitable positive integer depending on ω and ε.

Proof. . □

Let X, Y be real random variables on Ω.

Proposition 923 *Assume the sequence $(X_n)_{n\geq 1}$ converges almost surely to both X and Y. Then,*

$$X \overset{\mathbf{P}\text{-}\boldsymbol{a.s.}}{=} Y. \tag{11.1}$$

Proof. Since $(X_n)_{n\geq 1}$ converges almost surely to X [resp. Y] there exists an almost sure event $E \in \mathcal{E}$ [resp. $F \in \mathcal{E}$] such that

$$\lim_{n\to\infty} X_n(\omega) = X(\omega) \quad [\text{resp. } \lim_{n\to\infty} X_n(\omega) = Y(\omega)]$$

for every $\omega \in E$ [resp. $\omega \in F$]. Since $\mathbf{P}(F) = \mathbf{P}(E) = 1$, with regard to the event $E \cap F \in \mathcal{E}$ we also have

$$\mathbf{P}(E \cap F) = 1. \tag{11.2}$$

In fact,

$$\mathbf{P}(E \cap F) = \mathbf{P}(E \cup F) - (\mathbf{P}(E - F) + \mathbf{P}(F - E)),$$

where

$$\mathbf{P}(E - F) = \mathbf{P}(E \cap F^c) \leq \mathbf{P}(F^c) = 1 - \mathbf{P}(F) = 0,$$

$$\mathbf{P}(F-E)=\mathbf{P}(F\cap E^c)\leq\mathbf{P}(E^c)=1-\mathbf{P}(E)=0,$$

and

$$\mathbf{P}(E\cup F)\geq\max\{\mathbf{P}(E),\mathbf{P}(F)\}=1,$$

These imply Equation (11.2). Moreover, for any $\omega\in E\cap F$ and any $\varepsilon>0$, there exist $\ell(\omega,\varepsilon/2)$ and $m(\omega,\varepsilon/2)$ such that

$$|X_n(\omega)-X(\omega)|<\varepsilon/2\quad[\text{resp. }|X_n(\omega)-Y(\omega)|<\varepsilon/2] \tag{11.3}$$

for every $n\geq\ell(\omega,\varepsilon/2)$ [resp. $n\geq m(\omega,\varepsilon/2)$]. Setting

$$n(\omega,\varepsilon)\equiv\max\{\ell(\omega,\varepsilon/2),m(\omega,\varepsilon/2)\},$$

applying the triangular inequality on account of (11.3), we obtain

$$|X(\omega)-Y(\omega)|\leq|X_n(\omega)-X(\omega)|+|X_n(\omega)-Y(\omega)|<\varepsilon,$$

for every $n\geq n(\omega,\varepsilon)$. From the arbitrariness of ε it then follows

$$X(\omega)=Y(\omega),$$

for every $\omega\in E\cap F$, which is the desired result. □

Theorem 924 *The sequence $(X_n)_{n\geq1}$ converges almost surely to X, if and only if we have*

$$\lim_{m\to\infty}\mathbf{P}\left(\bigcap_{n\geq m}\{|X_n-X|<\varepsilon\}\right)=1, \tag{11.4}$$

or, equivalently,

$$\lim_{m\to\infty}\mathbf{P}\left(\bigcup_{n\geq m}\{|X_n-X|\geq\varepsilon\}\right)=0, \tag{11.5}$$

for every $\varepsilon>0$.

Proof. If $X_n\overset{\text{a.s.}}{\to}X$, then there exists an event $E\in\mathcal{E}$ such that $\mathbf{P}(E)=1$ and $\lim_{n\to\infty}X_n(\omega)=X(\omega)$, for every $\omega\in E$. This means that, for any $\omega\in E$ and any $\varepsilon>0$, there exists $m(\omega,\varepsilon)$ such that for every $n\geq m(\omega,\varepsilon)$ we have

$$|X_n(\omega)-X(\omega)|<\varepsilon.$$

Write

$$F_m(\varepsilon)\equiv\bigcap_{n\geq m}\{|X_n-X|<\varepsilon\},\quad\forall m\geq1.$$

Then, for any $\omega\in E$ there exists $m(\omega,\varepsilon)\in\mathbb{N}$ such that

$$\omega\in F_{m(\omega,\varepsilon)}(\varepsilon).$$

The latter implies

$$E\subseteq\bigcup_{m\geq1}F_m(\varepsilon).$$

On the other hand, the sequence $(F_m(\varepsilon))_{m\geq 1}$ is clearly non-decreasing. It follows,

$$\lim_{m\to\infty} \mathbf{P}(F_m(\varepsilon)) = \mathbf{P}\left(\bigcup_{m\geq 1} F_m(\varepsilon)\right) \geq \mathbf{P}(E) = 1.$$

Conversely, let us assume (11.4) holds true. Then, for any $\varepsilon > 0$ the event

$$F(\varepsilon) \equiv \bigcup_{m\geq 1} F_m(\varepsilon)$$

is almost sure. Thus, letting ε run through a sequence $(\delta_n)_{n\geq 1}$ of positive real numbers decreasing to zero, the event

$$E \equiv \bigcap_{n\geq 1} F(\delta_n)$$

is almost sure too. In fact, we have

$$\mathbf{P}(E) = \mathbf{P}\left(\bigcap_{n\geq 1} F(\delta_n)\right) = \lim_{n\to\infty} \mathbf{P}(F(\delta_n)) = 1.$$

Now, for any $\omega \in E$ we have $\omega \in F(\delta_n)$ for every $n \geq 1$. In particular, fixed any $\varepsilon > 0$, and choosing $n_\varepsilon \in \mathbb{N}$ such that $\delta_{n_\varepsilon} < \varepsilon$, we have

$$\omega \in F(\delta_{n_\varepsilon}) = \bigcup_{m\geq 1} F_m(\delta_{n_\varepsilon}).$$

Hence, for any $\omega \in E$ there exists $m(\omega, \varepsilon) \in \mathbb{N}$ such that

$$\omega \in F_{m(\omega,\varepsilon)}(\delta_{n_\varepsilon}) = \bigcap_{n\geq m(\omega,\varepsilon)} \{|X_n - X| < \delta_{n_\varepsilon}\}.$$

This implies

$$|X_n(\omega) - X(\omega)| < \varepsilon,$$

for every $n \geq m(\omega, \varepsilon)$. Thanks to the arbitrariness of ε, it Then, follows

$$\lim_{n\to\infty} X_n(\omega) = X(\omega),$$

which completes the proof. □

Corollary 925 *Assume we have*

$$\mathbf{P}\left(\limsup_{n\to\infty} \{|X_n - X| > \varepsilon\}\right) = 0, \tag{11.6}$$

for every $\varepsilon > 0$. Then, $X_n \overset{a.s.}{\to} X$. Conversely, assume we have

$$\mathbf{P}\left(\limsup_{n\to\infty} \{|X_n - X| > \varepsilon_0\}\right) > 0, \tag{11.7}$$

for some $\varepsilon_0 > 0$. Then, $X_n \overset{a.s.}{\not\to} X$.

Proof. Writing

$$F_m(\varepsilon) \equiv \bigcup_{n \geq m} \{|X_n - X| > \varepsilon\}, \quad \forall \varepsilon > 0,$$

we clearly have that the sequence $(F_m(\varepsilon))_{m \geq 1}$ is non-increasing. Hence, there exists

$$\lim_{m \to \infty} \mathbf{P}(F_m(\varepsilon)) = \mathbf{P}\left(\bigcap_{m \geq 1} F_m(\varepsilon)\right). \tag{11.8}$$

On the other hand,

$$\mathbf{P}\left(\bigcap_{m \geq 1} F_m(\varepsilon)\right) = \mathbf{P}\left(\bigcap_{m \geq 1} \bigcup_{n \geq m} \{|X_n - X| > \varepsilon\}\right) = \mathbf{P}\left(\limsup_{n \to \infty} \{|X_n - X| > \varepsilon\}\right).$$

Therefore, on account of (11.8), Equation (11.5) is satisfied or breached according to whether (11.6) or (11.7) holds true. Hence, applying Theorem 924, we obtain the desired result. □

Corollary 926 *Assume we have*

$$\sum_{n=1}^{\infty} \mathbf{P}(|X_n - X| > \varepsilon) < \infty, \quad \forall \varepsilon > 0. \tag{11.9}$$

Then, $X_n \overset{a.s.}{\to} X$.

Proof. Under (11.9), we have

$$\lim_{m \to \infty} \sum_{n=m}^{\infty} \mathbf{P}(|X_n - X| > \varepsilon) = 0, \quad \forall \varepsilon > 0. \tag{11.10}$$

On the other hand,

$$\mathbf{P}\left(\bigcup_{m \leq n} \{|X_n - X| > \varepsilon\}\right) \leq \sum_{n=m}^{\infty} \mathbf{P}(|X_n - X| > \varepsilon), \quad \forall \varepsilon > 0. \tag{11.11}$$

Combining (11.10) and (11.11), we then, obtain

$$\lim_{m \to \infty} \mathbf{P}\left(\bigcup_{m \leq n} \{|X_n - X| > \varepsilon\}\right) = 0,$$

which implies the desired claim in light of Theorem 924. □

Corollary 927 *Assume*

$$\sum_{n=1}^{\infty} \mathbf{P}(|X_n| > \varepsilon) < \infty, \quad \forall \varepsilon > 0. \tag{11.12}$$

Then, $X_n \overset{a.s.}{\to} 0$*, where* 0 *stands for the Dirac random variable on* Ω *concentrated at* 0.

Proof. . □

Corollary 928 *Assume there exists $\delta > 0$ such that*

$$\sum_{n=1}^{\infty} \mathbf{E}\left[|X_n|^{\delta}\right] < \infty. \tag{11.13}$$

Then, $X_n \overset{a.s.}{\to} 0$, where 0 stands for the Dirac random variable on Ω concentrated at 0.

Proof. Applying the Chebyshev inequality (5.195) with reference to the function $\varphi : \mathbb{R} \to \mathbb{R}$ given by

$$\varphi(x) \overset{\text{def}}{=} |x|^{\delta}, \quad \forall x \in \mathbb{R},$$

we have

$$\mathbf{P}(|X_n| > \varepsilon) \leq \frac{\mathbf{E}\left[|X_n|^{\delta}\right]}{\varepsilon^{\delta}},$$

for every $\varepsilon > 0$ and every $n \in \mathbb{N}$. Hence, (11.13) implies (11.12). The desired result immediately follows. □

Theorem 929 (Cauchy criterion) *The sequence $(X_n)_{n\geq 1}$ converges almost surely, if and only if we have*

$$\lim_{m\to\infty} \mathbf{P}\left(\bigcap_{m\leq n_1<n_2} \{|X_{n_1} - X_{n_2}| \leq \varepsilon\}\right) = 1, \quad \forall \varepsilon > 0 \tag{11.14}$$

or equivalently

$$\lim_{m\to\infty} \mathbf{P}\left(\bigcup_{m\leq n_1<n_2} \{|X_{n_1} - X_{n_2}| > \varepsilon\}\right) = 0, \quad \forall \varepsilon > 0. \tag{11.15}$$

Proof. . □

Theorem 930 *Assume the sequence $(X_n)_{n\geq 1}$ converges almost surely to X. Then, for every continuous function $g : \mathbb{R} \to \mathbb{R}$ we have $g(X_n) \overset{a.s.}{\to} g(X)$.*

Proof. Consider the events

$$E \equiv \left\{\lim_{n\to\infty} |X_n - X| = 0\right\}, \qquad F \equiv \left\{\lim_{n\to\infty} |g(X_n) - g(X)| = 0\right\}.$$

Since the function $g : \mathbb{R} \to \mathbb{R}$ is continuous, for any sample point $\omega \in E$ we have $\omega \in F$. This implies that, $E \subseteq F$. Now, since $X_n \overset{\text{a.s.}}{\to} X$ we have that $\mathbf{P}(E) = 1$. It Then, follows $\mathbf{P}(F) = 1$ which yields $g(X_n) \overset{\text{a.s.}}{\to} g(X)$, as claimed. □

Let $(X_n)_{n\geq 1}$ and $(Y_n)_{n\geq 1}$ [resp. X and Y] be sequences of real random variables [resp. real random variables] on Ω and let $\alpha, \beta \in \mathbb{R}$.

Theorem 931 *Assume $X_n \overset{a.s.}{\to} X$ and $Y_n \overset{a.s.}{\to} Y$. Then, we have:*

1. *$\alpha X_n + \beta Y_n \overset{a.s.}{\to} \alpha X + \beta Y$;*
2. *$X_n Y_n \overset{a.s.}{\to} XY$;*
3. *$X_n/Y_n \overset{a.s.}{\to} X/Y$, provided we have $\mathbf{P}(Y_n = 0) = 0$, definitively, and $\mathbf{P}(Y = 0) = 0$.*[1]*.*

Proof. . □

[1] From Example 918, it is seen that having $\mathbf{P}(Y_n = 0) = 0$ definitively does not imply that $\mathbf{P}(Y = 0) = 0$.

11.1.2 Convergence in Probability

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space and let $(X_n)_{n\geq 1}$ [resp. X] be a sequence of real random variables [resp. a real random variable] on Ω.

Definition 932 *We say that the sequence* $(X_n)_{n\geq 1}$ converges in probability *to* X *(as* n *goes to infinity), and we write*

$$X_n \xrightarrow{\boldsymbol{P}} X,$$

if we have

$$\lim_{n\to\infty} \mathbf{P}\left(|X_n - X| < \varepsilon\right) = 1 \tag{11.16}$$

or, equivalently,

$$\lim_{n\to\infty} \mathbf{P}\left(|X_n - X| \geq \varepsilon\right) = 0, \tag{11.17}$$

for every $\varepsilon > 0$.

Example 933 *Let* $(X_n)_{n\geq 1}$ *be a sequence of Bernoulli random variables on some probability space* Ω *such that*

$$X_n \overset{def}{=} \begin{cases} 1, & \mathbf{P}\left(X_n = 1\right) = \frac{1}{n}, \\ 0, & \mathbf{P}\left(X_n = 0\right) = 1 - \frac{1}{n}. \end{cases}$$

We have $X_n \xrightarrow{\boldsymbol{a.s.}} 0$, *where* 0 *stands for the Dirac random variable on* Ω *concentrated at* 0.

Discussion. We have

$$\mathbf{P}\left(|X_n - 0| \geq \varepsilon\right) = \mathbf{P}\left(X_n = 1\right) = \frac{1}{n},$$

for every $\varepsilon > 0$. Therefore, (11.17) of Definition 932 is clearly satisfied. □

Example 934 *Let* $(X_n)_{n\geq 1}$ *be a sequence of absolutely continuous real random variables on some probability space* Ω, *such that each* X_n *has density* $f_{X_n} : \mathbb{R} \to \mathbb{R}$ *given by*

$$f_{X_n}(x) \overset{def}{=} \frac{n}{\pi} \frac{1}{1 + n^2 x^2}, \quad \forall x \in \mathbb{R}, \quad \forall n \in \mathbb{N}.$$

Then, we have $X_n \xrightarrow{\boldsymbol{P}} 0$, *where* 0 *stands for the Dirac random variable on* Ω *concentrated at* 0.

Proof. Fixed any $\varepsilon > 0$ and $n \in \mathbb{N}$, we have

$$\begin{aligned} \mathbf{P}\left(|X_n| < \varepsilon\right) = \mathbf{P}\left(-\varepsilon < X_n < \varepsilon\right) &= \int_{-\varepsilon}^{\varepsilon} \frac{n}{\pi} \frac{1}{1 + n^2 x^2} dx = \frac{2}{\pi} \int_0^{\varepsilon} \frac{n}{1 + n^2 x^2} dx \\ &= \frac{2}{\pi} \int_0^{\varepsilon n} \frac{1}{1 + y^2} dy = \frac{2}{\pi} \arctan\left(y\right)\Big|_0^{\varepsilon n} = \frac{2}{\pi} \arctan\left(\varepsilon n\right). \end{aligned}$$

Therefore,

$$\lim_{n\to\infty} \mathbf{P}\left(|X_n| < \varepsilon\right) = \frac{2}{\pi} \lim_{n\to\infty} \arctan\left(\varepsilon n\right) = \frac{2}{\pi} \frac{\pi}{2} = 1,$$

for every $\varepsilon > 0$, and (11.16) is satisfied. □

Remark 935 *The sequence $(X_n)_{n\geq 1}$ converges in probability to X, if and only if the sequence $(\tilde{X}_n)_{n\geq 1}$ given by*

$$\tilde{X}_n \overset{def}{=} X_n - X, \quad \forall n \geq 1.$$

converges in probability to zero, to say $(\tilde{X}_n)_{n\geq 1}$ converges almost surely to the Dirac random variable concentrated at 0. In symbols,

$$\tilde{X}_n \xrightarrow{\boldsymbol{P}} 0,$$

where 0 stands for the Dirac random variable on Ω concentrated at 0.

Theorem 936 (Cauchy criterion) *The sequence $(X_n)_{n\geq 1}$ converges in probability, if and only if, for any $\varepsilon > 0$ and any $\delta > 0$, exists $n(\delta, \varepsilon) \in \mathbb{N}$ such that we have*

$$1 - \mathbf{P}\{|X_n - X_m| \leq \varepsilon\} < \delta,$$

or, equivalently,

$$\mathbf{P}\{|X_n - X_m| > \varepsilon\} < \delta.$$

for all $n > m \geq n(\omega, \varepsilon)$.

Proof. . □

Let X, Y be real random variables on Ω.

Proposition 937 *Assume the sequence $(X_n)_{n\geq 1}$ converges in probability to both X and Y. Then,*

$$X \overset{\mathbf{P}\text{-}\boldsymbol{a.s.}}{=} Y.$$

Proof. First, observe that we have

$$\mathbf{P}\{|X - Y| \geq \varepsilon\} = 0, \tag{11.18}$$

for every $\varepsilon > 0$. In fact, choosing any $n \geq 1$, thanks to the triangular inequality, we have

$$|X(\omega) - Y(\omega)| \leq |X(\omega) - X_n(\omega)| + |Y(\omega) - X_n(\omega)|,$$

for every $\omega \in \Omega$. Hence, $|X(\omega) - Y(\omega)| \geq \varepsilon$ clearly implies

$$|X(\omega) - X_n(\omega)| \geq \varepsilon/2 \quad \text{or} \quad |Y(\omega) - X_n(\omega)| \geq \varepsilon/2.$$

It fhen, follows

$$\{|X - Y| \geq \varepsilon\} \subseteq \{|X - X_n| \geq \varepsilon/2\} \cup \{|Y - X_n| \geq \varepsilon/2\}.$$

Therefore, we can write

$$\mathbf{P}(|X - Y| \geq \varepsilon) \leq \mathbf{P}(|X - X_n| \geq \varepsilon/2) + \mathbf{P}(|Y - X_n| \geq \varepsilon/2),$$

and, letting n go to infinity (11.18) follows. Now, choosing any sequence $(\varepsilon_n)_{n\geq 1}$ which converges to zero, we have

$$\{|X - Y| \neq 0\} \subseteq \bigcup_{n\geq 1} \{|X - Y| \geq \varepsilon_n\}.$$

Accordingly

$$\mathbf{P}(|X - Y| \neq 0) \leq \mathbf{P}\left(\bigcup_{n\geq 1} \{|X - Y| \geq \varepsilon_n\}\right) \leq \sum_{n=1}^{\infty} \mathbf{P}(|X - Y| \geq \varepsilon_n) = 0,$$

which is the desired result. □

Theorem 938 *Assume the sequence $(X_n)_{n\geq 1}$ converges almost surely to X. Then $(X_n)_{n\geq 1}$ converges in probability to X. In symbols,*

$$X_n \overset{a.s.}{\to} X \Rightarrow X_n \overset{P}{\to} X.$$

Proof. Under the assumption of almost sure convergence, we have

$$\lim_{m\to\infty} \mathbf{P}\left(\bigcup_{n\geq m} \{|X_n - X| > \varepsilon\}\right) = 0,$$

for every $\varepsilon > 0$.This, a fortiori, implies

$$\lim_{n\to\infty} \mathbf{P}\left(|X_n - X| > \varepsilon\right) = 0,$$

for every $\varepsilon > 0$, which is the desired result. □

Remark 939 *Assume the sequence $(X_n)_{n\geq 1}$ converges in probability to X. Then, in general, $(X_n)_{n\geq 1}$ does not converge almost surely to X. In symbols,*

$$X_n \overset{a.s.}{\to} X \not\Rightarrow X_n \overset{P}{\to} X.$$

We display the content of Remark 939 trough some examples.

Example 940 *With reference to Example 933, assume the random variables in the sequence $(X_n)_{n\geq 1}$ are independendent. Then, we still have $X_n \overset{P}{\to} 0$, but $X_n \overset{a.s.}{\not\to} 0$, where 0 stands for the Dirac random variable on Ω concentrated at 0.*

Discussion. Clearly, we still have $X_n \overset{\mathbf{P}}{\to} 0$. Now, if $(X_n)_{n\geq 1}$ converged almost surely, we would necessarily have $X_n \overset{\mathbf{a.s.}}{\to} 0$. On the hother hand, by virtue of the independence of the random variable in the sequence $(X_n)_{n\geq 1}$, choosing any $\varepsilon < 1$, we can write

$$\begin{aligned}\mathbf{P}\left(\bigcap_{n\geq m}\{|X_n - 0| \leq \varepsilon\}\right) &= \mathbf{P}\left(\bigcap_{n\geq m}\{X_n \leq \varepsilon\}\right) \leq \mathbf{P}\left(\bigcap_{n\geq m}^{2m}\{X_n \leq \varepsilon\}\right) = \prod_{n=m}^{2m}\mathbf{P}\left(X_n \leq \varepsilon\right) \\ &= \prod_{n=m}^{2m}\mathbf{P}\left(X_n = 0\right) = \prod_{n=m}^{2m}\left(1 - \frac{1}{n}\right) \leq \left(1 - \frac{1}{2m}\right)^m.\end{aligned}$$

As a consequence,

$$\lim_{m\to\infty} \mathbf{P}\left(\bigcap_{n\geq m}\{|X_n| \leq \varepsilon\}\right) \leq \lim_{m\to\infty}\left(1 - \frac{1}{2m}\right)^m = \lim_{m\to\infty}\left(\left(1 - \frac{1}{2m}\right)^{2m}\right)^{1/2} = \frac{1}{\sqrt{2}} < 1.$$

This prevents (11.4) to be be satisfied, that is to say $X_n \overset{\mathbf{a.s.}}{\not\to} 0$. □

Example 941 *With reference to Example 934, assume the random variables in the sequence $(X_n)_{n\geq 1}$ are independendent. Then, we still have $X_n \overset{P}{\to} 0$, but $X_n \overset{a.s.}{\not\to} 0$, , where 0 stands for the Dirac random variable on Ω concentrated at 0.*

Discussion. Clearly, we still have $X_n \xrightarrow{\mathbf{P}} 0$. As in Example 940, if $(X_n)_{n\geq 1}$ converged almost surely, we would necessarily have $X_n \xrightarrow{\mathbf{a.s.}} 0$. Again, by virtue of the independence of the random variable in the sequence $(X_n)_{n\geq 1}$ we can write

$$\begin{aligned}
\mathbf{P}\left(\bigcap_{n\geq m}\{|X_n-0|\leq\varepsilon\}\right) &= \mathbf{P}\left(\bigcap_{n\geq m}\{|X_n|\leq\varepsilon\}\right) \leq \mathbf{P}\left(\bigcap_{n\geq m}^{2m}\{|X_n|\leq\varepsilon\}\right)\\
&= \prod_{n=m}^{2m}\mathbf{P}\left(-\varepsilon<X_n<\varepsilon\right) = \prod_{n=m}^{2m}\int_{-\varepsilon}^{\varepsilon}\frac{n}{\pi}\frac{1}{1+n^2x^2}dx\\
&= \prod_{n=m}^{2m}\frac{2}{\pi}\int_0^{\varepsilon}\frac{n}{1+n^2x^2}dx = \prod_{n=m}^{2m}\frac{2}{\pi}\int_0^{\varepsilon n}\frac{1}{1+y^2}dy\\
&= \prod_{n=m}^{2m}\frac{2}{\pi}\arctan\left(y\right)\Big|_0^{\varepsilon n} = \prod_{n=m}^{2m}\frac{2}{\pi}\arctan\left(\varepsilon n\right)\\
&\leq \prod_{n=m}^{2m}\frac{2}{\pi}\arctan\left(2\varepsilon m\right) = \left(\frac{2}{\pi}\arctan\left(2\varepsilon m\right)\right)^m.
\end{aligned}$$

On the other hand, since the sequence $\left(\left(\frac{2}{\pi}\arctan\left(2\varepsilon m\right)\right)^m\right)_{m=1}^{\infty}$ is increasing on increasing of m and the exponential function is continuous, applying L'Hôpital's rule for the indeterminate form $\frac{0}{0}$, we can write

$$\begin{aligned}
&\lim_{m\to\infty}\left(\frac{2}{\pi}\arctan\left(2\varepsilon m\right)\right)^m\\
&= \lim_{x\to+\infty}\left(\frac{2}{\pi}\arctan\left(2\varepsilon x\right)\right)^x = \lim_{x\to+\infty}\exp\left(x\log\left(\frac{2}{\pi}\arctan\left(2\varepsilon x\right)\right)\right)\\
&= \exp\left(\lim_{x\to+\infty}x\log\left(\frac{2}{\pi}\arctan\left(2\varepsilon x\right)\right)\right) = \exp\left(\lim_{x\to+\infty}\frac{\log\left(\frac{2}{\pi}\arctan\left(2\varepsilon x\right)\right)}{\frac{1}{x}}\right)\\
&= \exp\left(\lim_{x\to+\infty}\frac{\frac{d}{dx}\log\left(\frac{2}{\pi}\arctan\left(2\varepsilon x\right)\right)}{\frac{d}{dx}\frac{1}{x}}\right) = \exp\left(\lim_{x\to+\infty}\frac{\frac{2\varepsilon}{(\arctan 2x\varepsilon)(1+4\varepsilon^2x^2)}}{-\frac{1}{x^2}}\right)\\
&= \exp\left(-\lim_{x\to+\infty}\frac{2\varepsilon x^2}{(\arctan 2x\varepsilon)\left(1+4\varepsilon^2x^2\right)}\right) = \exp\left(-\lim_{x\to+\infty}\frac{\varepsilon}{(\arctan 2x\varepsilon)\left(\frac{1}{2x^2}+2\varepsilon^2\right)}\right)\\
&= \exp\left(-\frac{\varepsilon}{\frac{\pi}{2}2\varepsilon^2}\right) = \exp\left(-\frac{1}{\pi\varepsilon}\right)
\end{aligned}$$

Therefore, we obtain

$$\lim_{m\to\infty}\mathbf{P}\left(\bigcap_{n\geq m}\{|X_n-0|\leq\varepsilon\}\right)\leq\exp\left(-\frac{1}{\pi\varepsilon}\right)<1,$$

for every $\varepsilon<1$. This prevents (11.4) to be be satisfied, that is to say $X_n \xrightarrow{\mathbf{a.s.}} 0$. □

Example 942 *With reference to the setting of Example 918, consider the indicator functions* $1_{[(j-1)/k,j/k]}$*, for any* $k\in\mathbb{N}$ *and any* $j=1,\ldots,k$*. Write* $f_{k,j}\equiv 1_{[(j-1)/k,j/k]}$ *and order these functions lexicographically, first according to* k *increasing and then, for each* k*, according to* j *increasing. Write* $\check{k}(n)$ *for the smallest positive integer such that* $k(k+1)/2\geq n$*, in symbols*

$$\check{k}(n)\equiv\arg\min_{k\in\mathbb{N}}\left\{k(k+1)/2\geq n\right\},$$

and write $(X_n)_{n\geq 1}$ *for the sequence of random variables given by*

$$X_n \overset{def}{=} f_{\check{k}(n),n-\check{k}(n)\left(\check{k}(n)-1\right)/2}, \quad \forall n \in \mathbb{N}.$$

Then, we have $X_n \overset{\boldsymbol{P}}{\to} 0$, *but* $X_n \overset{\boldsymbol{a.s.}}{\not\to} 0$, *where* 0 *stands for the Dirac random variable on* Ω *concentrated at* 0.

Discussion. Note that

$$\begin{aligned}
\check{k}(1) &= \arg\min_{k\in\mathbb{N}} \{k(k+1)/2 \geq 1\} = 1 \Rightarrow X_1 \equiv f_{\check{k}(1),1-\check{k}(1)\left(\check{k}(1)-1\right)/2} \equiv f_{1,1} \equiv 1_{[0,1]},\\
\check{k}(2) &= \arg\min_{k\in\mathbb{N}} \{k(k+1)/2 \geq 2\} = 2 \Rightarrow X_2 \equiv f_{\check{k}(1),2-\check{k}(1)\left(\check{k}(1)-1\right)/2} \equiv f_{2,1} \equiv 1_{[0,1/2]},\\
\check{k}(3) &= \arg\min_{k\in\mathbb{N}} \{k(k+1)/2 \geq 3\} = 2 \Rightarrow X_3 \equiv f_{\check{k}(1),3-\check{k}(1)\left(\check{k}(1)-1\right)/2} \equiv f_{2,2} \equiv 1_{[1/2,1]},\\
\check{k}(4) &= \arg\min_{k\in\mathbb{N}} \{k(k+1)/2 \geq 4\} = 3 \Rightarrow X_4 \equiv f_{\check{k}(1),4-\check{k}(1)\left(\check{k}(1)-1\right)/2} \equiv f_{3,1} \equiv 1_{[0,1/3]},\\
\check{k}(5) &= \arg\min_{k\in\mathbb{N}} \{k(k+1)/2 \geq 5\} = 3 \Rightarrow X_5 \equiv f_{\check{k}(1),5-\check{k}(1)\left(\check{k}(1)-1\right)/2} \equiv f_{3,2} \equiv 1_{[1/3,2/3]},\\
\check{k}(6) &= \arg\min_{k\in\mathbb{N}} \{k(k+1)/2 \geq 6\} = 3 \Rightarrow X_6 \equiv f_{\check{k}(1),6-\check{k}(1)\left(\check{k}(1)-1\right)/2} \equiv f_{3,3} \equiv 1_{[2/3,1]},\\
&\ldots
\end{aligned}$$

As a consequence,

$$\mathbf{P}(|X_n| > 0) = 1/\check{k}(n).$$

It clearly follows that $X_n \overset{\mathbf{P}}{\to} 0$. On the other hand, it is also evident that for every $\omega \in [0,1]$ there exist infinitely many values of n such that $X_n(\omega) = 1$ and infinitely many values of n such that $X_n(\omega) = 0$. Therefore, the subset of $[0,1]$ in which the sequence $(X_n)_{n\geq 1}$ converges is empty. □

Theorem 943 *If* $X_n \overset{\boldsymbol{P}}{\to} X$, *Then, there exists a subsequence* $(X_{n_k})_{k\geq 1}$ *of* $(X_n)_{n\geq 1}$ *such that* $X_{n_k} \overset{\boldsymbol{a.s.}}{\to} X$.

Proof. Considering a decreasing sequence of positive real numbers $(\varepsilon_m)_{m\geq 1}$ such that $\lim_{m\to\infty} \varepsilon_m = 0$, under the assumption of the Theorem for each ε_m we can determine a subsequence $(X_{n_k}(\varepsilon_m))_{k\geq 1}$ of $(X_n)_{n\geq 1}$ such that

$$\mathbf{P}(|X_{n_k}(\varepsilon_m) - X| > \varepsilon_m) < \frac{1}{2^k},$$

for every $k \geq 1$. We then, have

$$\sum_{k=1}^{\infty} \mathbf{P}(|X_{n_k}(\varepsilon_m) - X| > \varepsilon_m) \leq \sum_{k=1}^{\infty} \frac{1}{2^k} = 1.$$

Now, consider the sequence $(X_{n_m}(\varepsilon_m))_{m\geq 1}$ built from the set of sequences $\left\{(X_{n_k}(\varepsilon_m))_{k\geq 1},\ m \geq 1\right\}$ by the diagonal method. For any $\varepsilon > 0$ there exists $m(\varepsilon)$ such that $\varepsilon_m < \varepsilon$ for every $m > m(\varepsilon)$. It follows

$$\mathbf{P}(|X_{n_m}(\varepsilon_m) - X| > \varepsilon) \leq \mathbf{P}(|X_{n_m}(\varepsilon_m) - X| > \varepsilon_m) < \frac{1}{2^m},$$

for every $m > m(\varepsilon)$. Therefore,

$$\begin{aligned}
&\sum_{m=1}^{\infty} \mathbf{P}\left(|X_{n_m}(\varepsilon_m) - X| > \varepsilon\right) \\
&= \sum_{m=1}^{m(\varepsilon)} \mathbf{P}\left(|X_{n_m}(\varepsilon_m) - X| > \varepsilon\right) + \sum_{m=m(\varepsilon)+1}^{\infty} \mathbf{P}\left(|X_{n_m}(\varepsilon_m) - X| > \varepsilon\right) \\
&\leq \sum_{m=1}^{m(\varepsilon)} \mathbf{P}\left(|X_{n_m}(\varepsilon_m) - X| > \varepsilon\right) + \sum_{m=m(\varepsilon)+1}^{\infty} \frac{1}{2^m} \\
&< \infty
\end{aligned}$$

By virtue of the arbitrariness of $\varepsilon > 0$, we have that (11.9) holds true and Corollary 926 gives the desired result. □

Theorem 944 (Slutsky) *Assume $X_n \xrightarrow{\boldsymbol{P}} X$. Then, for every continuous function $g : \mathbb{R} \to \mathbb{R}$ we have $g(X_n) \xrightarrow{\boldsymbol{P}} g(X)$.*

Proof. . □

Let $(X_n)_{n\geq 1}$ and $(Y_n)_{n\geq 1}$ [resp. X and Y] be sequences of real random variables [resp. real random variables] on Ω and let $\alpha, \beta \in \mathbb{R}$.

Theorem 945 *Assume $X_n \xrightarrow{\boldsymbol{P}} X$ and $Y_n \xrightarrow{\boldsymbol{P}} Y$. Then, we have*

1. *$\alpha X_n + \beta Y_n \xrightarrow{\boldsymbol{P}} \alpha X + \beta Y$;*
2. *$X_n Y_n \xrightarrow{\boldsymbol{P}} XY$;*
3. *$X_n / Y_n \xrightarrow{\boldsymbol{P}} X/Y$, provided $\mathbf{P}(Y_n = 0)$, definitively, and $\mathbf{P}(Y = 0) = 0$.*

Proof. . □

11.1.3 Weak Convergence

Let $(X_n)_{n\geq 1}$ be a sequence of real random variables each of which defined on a probabiliy space $(\Omega_n, \mathcal{E}_n, \mathbf{P}_n) \equiv \Omega_n$, for every $n \in \mathbb{N}$, and let X be a real random variable on a probability space $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$.

Definition 946 *We say that the sequence* $(X_n)_{n\geq 1}$ converges weakly *or* in distribution *or* in law *to* X*, as* n *goes to infinity, and we write*

$$X_n \xrightarrow{\boldsymbol{w}} X \quad or \quad X_n \xrightarrow{\boldsymbol{d}} X \quad or \quad X_n \xrightarrow{\mathcal{L}} X,$$

if considering the sequence of distribution functions $(F_{X_n})_{n\geq 1}$ *corresponding to the sequence of randon variables* $(X_n)_{n\geq 1}$ *and the distribution function* F_X *of* X *we have*

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x),$$

for every $x \in \mathbb{R}$ *such that* F_X *is continuous at* x*.*

Example 947 *Consider the probability space* $(\Omega_n, \mathcal{E}_n, \mathbf{P}_n) \equiv \Omega_n$*, where*

$$\Omega_n \equiv \{1, \ldots, n\}, \qquad \mathcal{E}_n \equiv \mathcal{P}(\Omega_n), \qquad \mathbf{P}_n(E) \overset{def}{=} \frac{|E|}{n}, \quad \forall E \in \mathcal{E}_n, \quad \forall n \in \mathbb{N},$$

and consider the sequence $(X_n)_{n\geq 1}$ *of discrete and uniformly distributed real random variables* $X_n : \Omega_n \to \mathbb{R}$*, given by*

$$X_n(\omega_{n,k}) \overset{def}{=} \frac{k}{n}, \quad k = 1, \ldots, n, \quad \mathbf{P}_n\left(X_n = \frac{k}{n}\right) \overset{def}{=} \frac{1}{n}, \quad \forall n \in \mathbb{N}$$

Let $X \sim Unif(0,1)$*. Then,* $X_n \xrightarrow{\boldsymbol{w}} X$*.*

Discussion. Write $F_{X_n} : \mathbb{R} \to \mathbb{R}$ be the distribution function of X_n and let $F_X : \mathbb{R} \to \mathbb{R}$ be the distribution function of X. We have

$$F_{X_n}(x) = \begin{cases} 0, & \text{if } x < 1/n, \\ \frac{k}{n}, & \text{if } \frac{k}{n} \leq x < \frac{k+1}{n}, \quad k = 1, \ldots, n-1, \\ 1, & \text{if } x \geq 1. \end{cases}$$

and

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

It is evident that

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x), \quad \forall x \in (-\infty, 0] \cup [1, +\infty).$$

We prove that we have also

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x), \quad \forall x \in (0,1). \tag{11.19}$$

Observe that, for any $x \in (0,1)$ there exist $n(x) \in \mathbb{N}$ and $k(x,n) \in \{1,\ldots,n-1\}$ such that $x \in \left[\frac{k(x,n)}{n}, \frac{k(x,n)+1}{n}\right)$, for every $n > n(x)$. This because the familly of closed-open intervals $\left\{\left[\frac{k}{n}, \frac{k+1}{n}\right)\right\}_{k=1}^{n}$ is a partition of $\left[\frac{1}{n}, 1\right)$. We then have

$$\frac{k(x,n)}{n} = F_X\left(\frac{k(x,n)}{n}\right) \leq F_X(x) < F_X\left(\frac{k(x,n)+1}{n}\right) = \frac{k(x,n)+1}{n}$$

and

$$F_{X_n}(x) = \frac{k(x,n)}{n}.$$

As a consequence,

$$0 \leq F_X(x) - F_{X_n}(x) < \frac{k(x,n)+1}{n} - \frac{k(x,n)}{n} = \frac{1}{n},$$

for every $n > n(x)$. The latter yields the desired (11.19), taking the limit as $n \to \infty$. □

Let $(P_{X_n})_{n\geq 1}$ the sequence of distributions corresponding to the sequence of randon variables $(X_n)_{n\geq 1}$ and let P_X be the distribution of X. Recall that $P_{X_n} : \mathcal{B}(\mathbb{R}) \to \mathbb{R}$ [resp. $P_X : \mathcal{B}(\mathbb{R}) \to \mathbb{R}$] is given by

$$P_{X_n}(B) \overset{\text{def}}{=} \mathbf{P}_n(X_n \in B) \quad [\text{resp. } P_X(B) = \mathbf{P}(X \in B)], \quad \forall B \in \mathcal{B}(\mathbb{R}).$$

Theorem 948 *We have $X_n \xrightarrow{w} X$ if and only if*

$$\lim_{n\to\infty} P_{X_n}(B) = P_X(B), \tag{11.20}$$

for every $B \in \mathcal{B}(\mathbb{R})$ such that $P_X(\partial B) = 0$, where ∂B is the boundary of B.

Proof. . □

Theorem 949 *We have $X_n \xrightarrow{w} X$ if and only if one of the following equivalent conditions holds true*

$$\liminf_{n\to\infty} P_{X_n}(O) \leq P_X(O), \tag{11.21}$$

for every $O \in \mathcal{O}(\mathbb{R})$, and

$$\limsup_{n\to\infty} P_{X_n}(C) \leq P_X(C), \tag{11.22}$$

for every $C \in \mathcal{C}(\mathbb{R})$, where $\mathcal{O}(\mathbb{R})$ and $\mathcal{C}(\mathbb{R})$ are the families of all open and closed subets of $\mathbb{R}$, respectively.

Proof. . □

Let $C_b(\mathbb{R})$ be the Banach space of all continuous bounded real functions on $\mathbb{R}$ equipped with the supremum norm $\|\cdot\|_\infty$, given by

$$\|f\|_\infty \overset{\text{def}}{=} \sup_{x\in\mathbb{R}} \{|f(x)|\}, \quad \forall f \in C_b(\mathbb{R}).$$

Theorem 950 *We have* $X_n \xrightarrow{w} X$ *if and only if*

$$\lim_{n\to\infty} \int_{\mathbb{R}} f \, dP_{X_n} = \int_{\mathbb{R}} f \, dP_X, \tag{11.23}$$

for every $f \in C_b(\mathbb{R})$.
Proof. . □

Let $Lip_b(\mathbb{R})$ be the subspace of the Banach space $C_b(\mathbb{R})$ of all Lipschitz functions.

Theorem 951 *We have* $X_n \xrightarrow{w} X$ *if and only if*

$$\lim_{n\to\infty} \int_{\mathbb{R}} f \, dP_{X_n} = \int_{\mathbb{R}} f \, dP_X, \tag{11.24}$$

for every $f \in Lip_b(\mathbb{R})$.
Proof. . □

Let $(F_{X_n})_{n\geq 1}$ [resp. $(\varphi_{X_n})_{n\geq 1}$] be the sequence of distribution [resp. characteristic] functions corresponding to the sequence of randon variables $(X_n)_{n\geq 1}$ and let F_X [resp. φ_X] be the distribution [resp. characteristic] function of X.

Theorem 952 (Lévy) *We have*

1. *If* $X_n \xrightarrow{w} X$, *Then,* $\varphi_{X_n}(t) \to \varphi_X(t)$, *for every* $t \in \mathbb{R}$;
2. *If* $\varphi_{X_n}(t) \to \varphi(t)$, *for every* $t \in \mathbb{R}$, *where* $\varphi : \mathbb{R} \to \mathbb{R}$ *is continuous at* $t = 0$, *then* $\varphi = \varphi_X$ *and* $X_n \xrightarrow{w} X$.

Proof. . □

Corollary 953 *If* $\varphi_{X_n}(t) \to \varphi_X(t)$, *for every* $t \in \mathbb{R}$, *then* $X_n \xrightarrow{w} X$.

Proof. . □

Theorem 954 *Assume the random variables of the sequence* $(X_n)_{n\geq 1}$ *are all defined on the same probability space* $\Omega \equiv (\Omega, \mathcal{E}, \mathbf{P})$, *where* X *is defined. Assume also* $X_n \xrightarrow{P} X$. *Then,* $X_n \xrightarrow{w} X$.

Proof. First, observe that for all real random variables Y, Z on Ω we have

$$\mathbf{P}(Y \leq y) \leq \mathbf{P}(Z \leq y + \varepsilon) + \mathbf{P}(|Z - Y| > \varepsilon), \quad \forall y \in \mathbb{R}, \ \forall \varepsilon > 0. \tag{11.25}$$

In fact,

$$\begin{aligned}
\mathbf{P}(Y \leq y) &= \mathbf{P}(Y \leq y, Z \leq y + \varepsilon) + \mathbf{P}(Y \leq y, Z > y + \varepsilon) \\
&\leq \mathbf{P}(Z \leq y + \varepsilon) + \mathbf{P}(Y - Z \leq y - Z, y - Z < -\varepsilon) \\
&\leq \mathbf{P}(Z \leq y + \varepsilon) + \mathbf{P}(Y - Z < -\varepsilon) \\
&\leq \mathbf{P}(Z \leq y + \varepsilon) + \mathbf{P}(Y - Z < -\varepsilon) + \mathbf{P}(Y - Z > \varepsilon) \\
&\leq \mathbf{P}(Z \leq y + \varepsilon) + \mathbf{P}(|Y - Z| > \varepsilon),
\end{aligned}$$

as desired. Now, consider any point x where the distribution function F_X is continuous and apply (11.25) by setting $Y \equiv X_n$, $Z \equiv X$, and with reference to the point $y \equiv x$. We have

$$\mathbf{P}(X_n \leq x) \leq \mathbf{P}(X \leq x + \varepsilon) + \mathbf{P}(|X_n - X| > \varepsilon) \tag{11.26}$$

In addition, apply (11.25) by setting $Y \equiv X$, $Z \equiv X_n$, and with reference to the point $y \equiv x - \varepsilon$. We have

$$\mathbf{P}(X \leq x - \varepsilon) \leq \mathbf{P}(X_n \leq x) + \mathbf{P}(|X_n - X| > \varepsilon),$$

that is to say

$$\mathbf{P}(X \leq x - \varepsilon) - \mathbf{P}(|X_n - X| > \varepsilon) \leq \mathbf{P}(X_n \leq x), \tag{11.27}$$

Combining (11.26) and (11.27), it then follows

$$\mathbf{P}(X \leq x - \varepsilon) - \mathbf{P}(|X_n - X| > \varepsilon) \leq \mathbf{P}(X_n \leq x) \leq \mathbf{P}(X \leq x + \varepsilon) + \mathbf{P}(|X_n - X| > \varepsilon).$$

Taking the limit as $n \to \infty$, under the assumption $X_n \xrightarrow{\mathbf{P}} X$, we obtain

$$\mathbf{P}(X \leq x - \varepsilon) \leq \lim_{n\to\infty} \mathbf{P}(X_n \leq x) \leq \mathbf{P}(X \leq x + \varepsilon).$$

That is

$$F_X(x - \varepsilon) \leq \lim_{n\to\infty} F_{X_n}(x) \leq F_X(x + \varepsilon). \tag{11.28}$$

Now, considering a point x where F_X is continuous and taking the limit as $\varepsilon \to 0^+$ the inequality chain (11.28) yields

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x),$$

that is the weak convergence of $(X_n)_{n\geq 1}$ to X. □

Remark 955 *In general, convergence in distribution does not imply convergence in probability.*

Example 956 *Let X be a symmetric real random variable, for instance $X \sim N(0,1)$, and let $(X_n)_{n\geq 0}$ be the sequence of real random variables given by*

$$X_n \overset{def}{=} -X, \quad \forall n \in \mathbb{N}.$$

Then, $X_n \xrightarrow{\boldsymbol{w}} X$ but, in general, $X_n \overset{\boldsymbol{P}}{\not\to} X$.

Discussion. Considering the distribution function $P_{X_n} : \mathcal{B}(\mathbb{R}) \to \mathbb{R}$ [resp. $P_X : \mathcal{B}(\mathbb{R}) \to \mathbb{R}$] of the random variable X_n [resp. X], we have

$$P_{X_n} = P_{-X} = P_X.$$

Therefore, $X_n \xrightarrow{\mathbf{w}} X$ (note that we have also $X_n \xrightarrow{\mathbf{w}} -X$). On the other hand, we clearly have $X_n \xrightarrow{\mathbf{a.s.}} -X$, which implies $X_n \xrightarrow{\mathbf{P}} -X$. Now, if we had also $X_n \xrightarrow{\mathbf{P}} X$, it would follow $X = -X$ a.s. on Ω, that is $X = 0$ a.s. on Ω, which is clearly false if $X \sim N(0,1)$. □

Proposition 957 *Assume the random variables of the sequence $(X_n)_{n\geq 1}$ are all defined on the same probability space $\Omega \equiv (\Omega, \mathcal{E}, \mathbf{P})$. Assume also that $X_n \xrightarrow{\boldsymbol{w}} Dirac(x_0)$ for some $x_0 \in \mathbb{R}$. Then, $X_n \xrightarrow{\boldsymbol{P}} Dirac(x_0)$.*

Proof. We have

$$\mathbf{P}\left(|X_n - x_0 1_\Omega| \leq \varepsilon\right) = \mathbf{P}\left(-(x_0+\varepsilon) \leq X_n \leq x_0+\varepsilon\right) = P_{X_n}\left(-(x_0+\varepsilon), x_0+\varepsilon\right)$$

On the other hand, under the assumption $X_n \xrightarrow{\mathbf{w}} x_0 \cdot 1_\Omega$, taking the limit as $n \to \infty$, we obtain

$$\lim_{n\to\infty} \mathbf{P}\left(|X_n - x_0 1_\Omega| \leq \varepsilon\right) = \lim_{n\to\infty} P_{X_n}\left(-(x_0+\varepsilon), x_0+\varepsilon\right) = P_{x_0 1_\Omega}\left(-(x_0+\varepsilon), x_0+\varepsilon\right) = 1,$$

because $P_{x_0 1_\Omega}$ is the Dirac distribution concentrated at x_0. This implies $X_n \xrightarrow{\mathbf{P}} Dirac\left(x_0\right)$. □

Theorem 958 *Assume $X_n \xrightarrow{w} X$. Then, for every continuous function $g : \mathbb{R} \to \mathbb{R}$ we have $g\left(X_n\right) \xrightarrow{w} g\left(X\right)$.*

Proof. . □

Let $\left(X_n\right)_{n\geq 1}$ and $\left(Y_n\right)_{n\geq 1}$ [resp. X and Y] be sequences of real random variables [resp. real random variables] on Ω and let $\alpha, \beta \in \mathbb{R}$.

Proposition 959 *Assume $X_n \xrightarrow{w} X$ and $Y_n \xrightarrow{w} Y$. In addition, assume X_n and Y_n are uncorrelated for every $n \in \mathbb{N}$. Then, $\alpha X_n + \beta Y_n \xrightarrow{w} \alpha X + \beta Y$.*

Proof. . □

Remark 960 *In general, the assumption that X_n and Y_n are uncorrelated for every $n \in \mathbb{N}$ cannot be dropped.*

Theorem 961 *Assume $X_n \xrightarrow{w} X$ and $Y_n \xrightarrow{P} Dirac\left(x_0\right)$ for some $x_0 \in \mathbb{R}$. Then, $X_n Y_n \xrightarrow{P} x_0 X$.*

Proof. . □

11.1.4 Convergence in pth-Mean

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a complete probability space. Fixed any $p \in \mathbb{R}_+$ such that $p \geq 1$, let $L^p(\Omega; \mathbb{R})$ be the Banach space of all equivalence classes of real random variables on Ω having finite moment of order p with respect to the $\mathbf{P}$-a.s. equalitity equivalence relationship. For our purposes we can neglect the difference between a real random variable with finite moment of order p and the equivalence class which the random variable belongs to. Thus, identifying the equivalence classes with the belonging random variables, we can think of $L^p(\Omega; \mathbb{R})$ as the Banach space of all real random variables on Ω having finite moment of order p. In this setting, consider a sequence $(X_n)_{n\geq 1}$ of real random variables in $L^p(\Omega; \mathbb{R})$ and let $X \in L^p(\Omega; \mathbb{R})$.

Definition 962 *We say that the sequence* $(X_n)_{n\geq 1}$ converges in pth-mean *to* X*, as* n *goes to infinity, and we write* $X_n \overset{\boldsymbol{L^p}}{\to} X$*, if we have*

$$\lim_{n\to\infty} \|X_n - X\|_p \equiv \lim_{n\to\infty} \mathbf{E}\left[|X_n - X|^p\right]^{1/p} = 0.$$

In particular, we say that the sequence $(X_n)_{n\geq 1}$ converges in mean *to* X*, as* n *goes to infinity, and we write* $X_n \overset{\boldsymbol{L^1}}{\to} X$*, if we have*

$$\lim_{n\to\infty} \|X_n - X\|_1 \equiv \lim_{n\to\infty} \mathbf{E}\left[|X_n - X|\right] = 0.$$

We say that the sequence $(X_n)_{n\geq 1}$ converges in mean square *to* X*, as* n *goes to infinity, and we write* $X_n \overset{\boldsymbol{L^2}}{\to} X$*, if we have*

$$\lim_{n\to\infty} \|X_n - X\|_2 \equiv \lim_{n\to\infty} \mathbf{E}\left[(X_n - X)^2\right]^{1/2} = 0.$$

Theorem 963 (Cauchy criterion) *The sequence* $(X_n)_{n\geq 1}$ *converges in* p*th-mean, if and only if for every* $\delta > 0$ *and for every* $\varepsilon > 0$ *there exists* $n(\delta, \varepsilon) \in \mathbb{N}$ *such that for all* $n > m \geq n(\omega, \varepsilon)$ *we have*

$$\mathbf{E}\left[|X_n - X_m|^p\right] < \delta,$$

Proposition 964 *Suppose that the sequence* $(X_n)_{n\geq 1}$ *converges in* p*th-mean to the real random variables* X *and* Y *on* Ω*. Then,*

$$X \overset{\mathbf{P}\text{-}\boldsymbol{a.s.}}{=} Y.$$

Theorem 965 *Assume* $X_n \overset{\boldsymbol{L^p}}{\to} X$*, for some* $p \geq 1$*. Then,* $X_n \overset{\boldsymbol{P}}{\to} X$*.*

Proof. Assume $X_n \overset{\mathbf{L}^P}{\to} X$, for some $p \geq 1$. Applying the first Markov inequality (see Corollary 651), we can write

$$\mathbf{P}\left(|X_n - X| \geq \varepsilon\right) = \mathbf{P}\left(|X_n - X|^p \geq \varepsilon^p\right) \leq \frac{\mathbf{E}\left[|X_n - X|^p\right]}{\varepsilon^p},$$

for every $\varepsilon > 0$. Therefore, since $\lim_{n\to\infty} \mathbf{E}\left[|X_n - X|^p\right] = 0$, we obtain $\lim_{n\to\infty} \mathbf{P}\left(|X_n - X| \geq \varepsilon\right) = 0$, for every $\varepsilon > 0$, which is the convergence in probability. □

Corollary 966 *Assume* $X_n \overset{\boldsymbol{L^p}}{\to} X$*, Then, there exists a subsequence* $(X_{n_k})_{k\geq 1}$ *of* $(X_n)_{n\geq 1}$ *which converges to* X *almost surely.*

Proof. Combining Theorem 965 and 943, the claim immediately follows. □

Remark 967 *In general, convergence in probability does not imply convergence in pth-mean.*

Theorem 968 *Assume $X_n \xrightarrow{\boldsymbol{P}} X$ and there exists $Y \in L^p(\Omega;\mathbb{R})$ such that $|X_n| \leq Y$ for every $n \geq 1$. Then, $X \in L^p(\Omega;\mathbb{R})$ and $X_n \xrightarrow{\boldsymbol{L^p}} X$.*

Proof. We can write

$$|X - X_n| \geq ||X| - |X_n|| \geq |X| - |X_n| \geq |X| - Y,$$

for every $n \geq 1$. Hence, fixed any $\varepsilon > 0$, we have

$$|X| - Y \geq \varepsilon \Rightarrow |X - X_n| \geq \varepsilon,$$

for every $n \geq 1$. This implies

$$\{|X| \geq Y + \varepsilon\} \subseteq \{|X - X_n| \geq \varepsilon\},$$

which in turn yields,

$$\mathbf{P}(|X| \geq Y + \varepsilon) \leq \mathbf{P}(|X - X_n| \geq \varepsilon)$$

independently of n. As a consequence, considering the limit as n goes to infinity, by virtue of the assumption $X_n \xrightarrow{\mathbf{P}} X$, we obtain

$$\mathbf{P}(|X| \geq Y + \varepsilon) \leq \lim_{n\to\infty} \mathbf{P}(|X - X_n| \geq \varepsilon) = 0.$$

On account of the Minkowski lemma (see Lemma **??**), it Then, follows,

$$\begin{aligned}\int_\Omega |X|^p \, d\mathbf{P} &= \int_{\{|X|\geq Y+\varepsilon\}} |X|^p \, d\mathbf{P} + \int_{\{|X|<Y+\varepsilon\}} |X|^p \, d\mathbf{P} \\ &\leq \int_{\{|X|<Y+\varepsilon\}} (Y+\varepsilon)^p \, d\mathbf{P} \leq \int_\Omega (Y+\varepsilon)^p \, d\mathbf{P} \\ &\leq 2^{p-1} \int_\Omega (Y^p + \varepsilon^p) \, d\mathbf{P} = 2^{p-1} \left(\int_\Omega Y^p d\mathbf{P} + \varepsilon^p \right) < \infty,\end{aligned}$$

which shows that $X \in L^p(\Omega;\mathbb{R})$. Now, assume that $(X_n)_{n\geq 1}$ does not converge to X in $L^p(\Omega;\mathbb{R})$. Then, it is possible to determine a subsequence $(X_{n_k})_{k\geq 1}$ such that

$$\mathbf{E}[|X - X_{n_k}|^p] \geq \varepsilon \tag{11.29}$$

for some $\varepsilon > 0$ and for every $k \geq 1$. Clearly, $X_{n_k} \xrightarrow{\mathbf{P}} X$. Therefore, considering Theorem 943, there exists a subsequence $\left(X_{n_{k_j}}\right)_{j\geq 1}$ of $(X_{n_k})_{k\geq 1}$ such that $X_{n_{k_j}} \xrightarrow{\mathbf{a.s.}} X$. Now, still by virtue of the Minkowski Lemma, we can write

$$\left|X - X_{n_{k_j}}\right|^p \leq 2^{p-1} \left(|X|^p + \left|X_{n_{k_J}}\right|^p \right) \leq 2^{p-1} \left(|X|^p + Y^p \right).$$

Hence, the sequence $\left(X - X_{n_{k_j}}\right)^p_{j\geq 1}$ is dominated in $L^1(\Omega;\mathbb{R})$ by the random variable $2^{p-1}(|X|^p + Y^p)$. Thanks to the Lebesgue Dominated Convergence Theorem we Then, obtain

$$\lim_{j\to\infty} \mathbf{E}\left[\left|X - X_{n_{k_j}}\right|^p\right] = 0,$$

which contradicts (11.29). This proves that $(X_n)_{n\geq 1}$ converges to X in $L^p(\Omega;\mathbb{R})$. □

Theorem 969 *Let $(X_n)_{n\geq 1}$ be a sequence of Gaussian random variables, $X_n \sim N(\mu_n, \sigma_n^2)$ for suitable $\mu_n, \sigma_n \in \mathbb{R}$ such that $\sigma_n > 0$, for every $n \in \mathbb{N}$. Then, $X_n \xrightarrow{\boldsymbol{P}} X$ implies $X_n \xrightarrow{\boldsymbol{L}^p} X$.*

Theorem 970 *Assume $X_n \xrightarrow{\boldsymbol{L}^p} X$, for some $p \geq 1$. Then, $X_n \xrightarrow{\boldsymbol{L}^\ell} X$, for every $\ell \in [1, p]$.*

Proof. By virtue of Lyapunov inequality (see Theorem 663), the assumption $X_n, X \in L^p(\Omega; \mathbb{R})$, for every $n \in \mathbb{N}$, implies $X_n, X \in L^\ell(\Omega; \mathbb{R})$, for every $\ell \in [1, p]$ and every $n \in \mathbb{N}$. In addition, we have

$$\|X_n - X\|_\ell \leq \|X_n - X\|_p,$$

for every $\ell \in [1, p]$ and every $n \in \mathbb{N}$. Therefore, the assumption $X_n \xrightarrow{\mathbf{L}^p} X$ clearly implies the desired result. □

Let $(X_n)_{n\geq 1}$ and $(Y_n)_{n\geq 1}$ be sequences of real random variables in $L^p(\Omega; \mathbb{R})$ let $X, Y \in L^p(\Omega; \mathbb{R})$ and let $\alpha, \beta \in \mathbb{R}$.

Theorem 971 *Assume $X_n \xrightarrow{\boldsymbol{L}^p} X$ and $Y_n \xrightarrow{\boldsymbol{L}^p} Y$. Then, $\alpha X_n + \beta Y_n \xrightarrow{\boldsymbol{L}^p} \alpha X + \beta Y$.*

Let $p, q \geq 1$ such that $1/p + 1/q = 1$. Let $(X_n)_{n\geq 1}$ [resp. $(Y_n)_{n\geq 1}$] be a sequence of real random variables in $L^p(\Omega; \mathbb{R})$ [resp. $L^q(\Omega; \mathbb{R})$] and let $X \in L^p(\Omega; \mathbb{R})$ [resp. $Y \in L^q(\Omega; \mathbb{R})$].

Theorem 972 *Assume $X_n \xrightarrow{\boldsymbol{L}^p} X$ and $Y_n \xrightarrow{\boldsymbol{L}^q} Y$. Then*

1. *$X_n \cdot Y_n \xrightarrow{\boldsymbol{L}^1} X \cdot Y$;*
2. *$X_n / Y_n \xrightarrow{\boldsymbol{L}^1} X/Y$, provided $\mathbf{P}(Y_n = 0)$, definitively, and $\mathbf{P}(Y = 0) = 0$.*

Convergence in mean

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space, let $L^1(\Omega; \mathbb{R})$ be the Banach space of all real random variables on Ω having finite moment of order 1, and let $(X_n)_{n\geq 1}$ be a sequence of real random variables in $L^1(\Omega; \mathbb{R})$.

Definition 973 *We say that $(X_n)_{n\geq 1}$ is uniformly integrable if*

$$\sup_{n\in\mathbb{N}} \{\mathbf{E}[|X_n|]\} < \infty, \tag{11.30}$$

and for every $\varepsilon > 0$ there exists $k > 0$ such that

$$\sup_{n\in\mathbb{N}} \left\{ \int_{\{|X_n|>k\}} |X_n| \, d\mathbf{P} \right\} < \varepsilon. \tag{11.31}$$

Remark 974 *The sequence $(X_n)_{n\geq 1}$ is uniformly integrable if and only if the sequence $(|X_n|)_{n\geq 1}$ is.*

Remark 975 *If there exists $X \in L^1(\Omega; \mathbb{R})$ such that*

$$|X_n| \leq X, \quad \forall n \in \mathbb{N},$$

Then, $(X_n)_{n\geq 1}$ is uniformly integrable.

Let $(X_n)_{n\geq 1}$ and $(Y_n)_{n\geq 1}$ be sequences of real random variables in $L^1(\Omega;\mathbb{R})$ and let $\alpha,\beta\in\mathbb{R}$.

Remark 976 *If $(X_n)_{n\geq 1}$ and $(Y_n)_{n\geq 1}$ are uniformly integrable Then, $(\alpha X_n+\beta Y_n)_{n\geq 1}$ is uniformly integrable.*

Theorem 977 *The following conditions are equivalent:*

1. *the sequence $(X_n)_{n\geq 1}$ is uniformly integrable;*
2. *we have $\lim_{k\to+\infty}\sup_{n\in\mathbb{N}}\int_{\{|X_n|>k\}}|X_n|\,d\mathbf{P}=0$;*
3. *there exists a convex function $\phi:\mathbb{R}\to\mathbb{R}$, with the properties $\phi(0)=0$, $\phi(-x)=\phi(x)$, and $\lim_{x\to+\infty}\phi(x)/x=+\infty$, such that $\sup_{n\in\mathbb{N}}\mathbf{E}[\phi(X_n)]<\infty$.*

Theorem 978 (Scheffe) *Assume the sequence $(X_n)_{n\geq 1}$ converges almost surely to $X\in L^1(\Omega;\mathbb{R})$. Then, $X_n\overset{\boldsymbol{L^1}}{\to}X$ if and only if $(X_n)_{n\geq 1}$ is uniformly integrable.*

Theorem 979 (Vitali) *Assume the sequence $(X_n)_{n\geq 1}$ converges almost surely to $X\in L^1(\Omega;\mathbb{R})$ and $(X_n)_{n\geq 1}$ is uniformly integrable. Then, we have*

$$\lim_{n\to\infty}\mathbf{E}[X_n]=\mathbf{E}[X].$$

Theorem 980 *Assume the sequence $(X_n)_{n\geq 1}$ converges in probability to $X\in L^1(\Omega;\mathbb{R})$ and $(X_n)_{n\geq 1}$ is uniformly integrable. Then, $X_n\overset{\boldsymbol{L^1}}{\to}X$. Conversely, assume that $X_n\overset{\boldsymbol{L^1}}{\to}X$ Then, $(X_n)_{n\geq 1}$ converges in probability to $X\in L^1(\Omega;\mathbb{R})$ and $(X_n)_{n\geq 1}$ is uniformly integrable.*

Remark 981 *In general, convergence in probability does not imply convergence in mean.*

Example 982 *Let $(x_n)_{n\geq 1}$ a sequence of positive real numbers and let $(X_n)_{n\geq 1}$ be a sequence of Bernoulli random variables such that*

$$\mathbf{P}(X_n=0)=1-\frac{1}{n},\quad \mathbf{P}(X_n=x_n)=\frac{1}{n}.$$

Then, $X_n\overset{\boldsymbol{P}}{\to}0$, but, depending on the choice of the sequence $(x_n)_{n\geq 1}$, it may not converge in mean.

Discussion. We have

$$\mathbf{P}(|X_n|\leq\varepsilon)\geq\mathbf{P}(X_n=0)=1-\frac{1}{n},$$

for every $\varepsilon>0$. Therefore,

$$\lim_{n\to\infty}\mathbf{P}(|X_n|\leq\varepsilon)=1,$$

which yields $X_n\overset{\mathbf{P}}{\to}0$. On the other hand we have

$$\mathbf{E}[|X_n|]=\mathbf{E}[X_n]=\frac{x_n}{n}.$$

Therefore, depending on the choice of the sequence $(x_n)_{n\geq 1}$ we may have

$$\lim_{n\to\infty}\mathbf{E}[|X_n|]\neq 0$$

which prevents the convergence in mean of $(X_n)_{n\geq 1}$. □

Convergence in mean square

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a complete probability space, let $L^2(\Omega; \mathbb{R})$ be the Hilbert space of all real random variables on Ω having finite moment of order 2, and let $(X_n)_{n\geq 1}$ be a sequence of real random variables in $L^2(\Omega; \mathbb{R})$.

Proposition 983 *Assume there exists $c \in \mathbb{R}$ such that for every $\varepsilon > 0$ it is possible to determine n_ε such that for every $m, n > n_\varepsilon$ we have*

$$|\mathbf{E}[X_m X_n] - c| < \varepsilon.$$

Then, there exists $X \in L^2(\Omega; \mathbb{R})$ such that $X_n \xrightarrow{L^2} X$.

Proposition 984 *Assume there exists $x_0 \in \mathbb{R}$ such that $\lim_{n\to\infty} \mathbf{E}[X_n] = x_0$ and $\lim_{n\to\infty} \mathbf{D}^2[X_n] = 0$. Then, we have $X_n \xrightarrow{L^2} x_0 1_\Omega$. Conversely, assume there exists $x_0 \in \mathbb{R}$ such that $X_n \xrightarrow{L^2} x_0 1_\Omega$. Then, we have $\lim_{n\to\infty} \mathbf{E}[X_n] = x_0$ and $\lim_{n\to\infty} \mathbf{D}^2[X_n] = 0$.*

Remark 985 *Almost sure convergence does not imply convergence in mean square.*

Example 986 *Let $(X_n)_{n\geq 1}$ be a sequence of random variables such that*

$$\mathbf{P}(X_n = -n) = \frac{1}{2n^\alpha}, \quad \mathbf{P}(X_n = 0) = 1 - \frac{1}{n^\alpha}, \quad \mathbf{P}(X_n = n) = \frac{1}{2n^\alpha}.$$

Then, $(X_n)_{n\geq 1}$ converges in probability, but not almost surely.

$$
\begin{array}{ccccccccc}
 & & X_n \overset{\mathbf{a.s.}}{\to} X & & \nRightarrow & & X_n \overset{\mathbf{L}^p}{\to} X & \Rightarrow & X_{n_k} \overset{\mathbf{a.s.}}{\to} X \\
 & & & & \nLeftarrow & & & & \\
 & & \Downarrow & & & & \Downarrow & \searrow & \\
X_{n_k} \overset{\mathbf{a.s.}}{\to} X & \Leftarrow & X_n \overset{\mathbf{P}}{\to} X & & X_n \overset{\mathbf{P}}{\to} X \text{ if } (\Omega_n = \Omega \text{ and } X \tilde{} Dir\,(x_0)) & & X_n \overset{\mathbf{P}}{\to} X & & X_n \overset{\mathbf{L}^q}{\to} X \text{ if } q \in [1, p \\
 & & \Downarrow & \nearrow & & & & & \\
 & & X_n \overset{\mathbf{w}}{\to} X & & & & & &
\end{array}
$$

11.2 Sequence of Independent and Identically Distributed Real Random Variables

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space, let $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu_L) \equiv \mathbb{R}$ be the real Euclidean state space, and let $(X_n)_{n\geq 1}$ be a sequence of real random variables on Ω.

Definition 987 *We say that $(X_n)_{n\geq 1}$ is a* first-order *[resp.* second-order*]* sequence *if all random variables of the sequence have finite moment of order* 1 *[resp.* 2*]. In symbols,*

$$X_n \in L^1(\Omega;\mathbb{R}) \quad \textit{[resp. } X_n \in L^2(\Omega;\mathbb{R})\textit{]}, \quad \forall n \geq 1.$$

Let $(X_n)_{n\geq 1}$ be a second-order sequence of real random variables on Ω.

Definition 988 *We say that $(X_n)_{n\geq 1}$ is a sequence of* uncorrelated *real random variables if the random variables of the sequence are pairwise uncorrelated. In symbols,*

$$\mathbf{E}[X_m X_n] = \mathbf{E}[X_m]\,\mathbf{E}[X_n], \quad \forall m, n \geq 1,\ m \neq n. \tag{11.32}$$

Let $(X_n)_{n\geq 1}$ be a second order sequence of uncorrelated real random variables on Ω.

Proposition 989 *We have*

$$\mathbf{D}^2\left[\sum_{k=1}^{n} X_k\right] = \sum_{k=1}^{n} \mathbf{D}^2[X_k],$$

for every $n \geq 1$.

Proof. A straightforward computation gives

$$\begin{aligned}
\mathbf{D}^2\left[\sum_{k=1}^{n} X_k\right] &= \mathbf{E}\left[\left(\sum_{k=1}^{n} X_k\right)^2\right] - \mathbf{E}\left[\sum_{k=1}^{n} X_k\right]^2 \\
&= \mathbf{E}\left[\sum_{j,k=1}^{n} X_j X_k\right] - \left(\sum_{k=1}^{n} \mathbf{E}[X_k]\right)^2 \\
&= \sum_{j,k=1}^{n} \mathbf{E}[X_j X_k] - \sum_{j,k=1}^{n} \mathbf{E}[X_j]\,\mathbf{E}[X_k]^{\intercal} \\
&= \sum_{k=1}^{n} \mathbf{E}\left[X_k^2\right] + \sum_{\substack{j,k=1 \\ j\neq k}}^{n} \mathbf{E}[X_j X_k] - \sum_{k=1}^{n} \mathbf{E}[X_k]^2 - \sum_{\substack{j,k=1 \\ j\neq k}}^{n} \mathbf{E}[X_j]\,\mathbf{E}[X_k],
\end{aligned}$$

for every $n \geq 1$. On account of (11.32), it Then, follows

$$\mathbf{D}^2\left[\sum_{k=1}^{n} X_k\right] = \sum_{k=1}^{n} \left(\mathbf{E}\left[X_k^2\right] - \mathbf{E}[X_k]^2\right),$$

for every $n \geq 1$, which is the desired result. □

Definition 990 *We say that* $(X_n)_{n\geq 1}$ *is a sequence of* independent *real random variables if the random variables of the sequence are totally independent. That is,*

$$\mathbf{P}\left(X_{n_1} \in B_1, \ldots, X_{n_m} \in B_m\right) = \prod_{j=1}^{m} \mathbf{P}\left(X_{n_j} \in B_j\right),$$

for any finite sequence $(n_j)_{j=1}^m$ *in* $\mathbb{N}$*, and any finite sequence* $(B_j)_{j=1}^m$ *in* $\mathcal{B}(\mathbb{R})$*.*

From Proposition 753, it immediately follows that second-order sequence of independent random variables is a sequence of uncorrelated random variables.

Let $(X_n)_{n\geq 1}$ be a sequence of real random variables on Ω.

Definition 991 *We call the* random walk *generated by* $(X_n)_{n\geq 1}$ *the sequence* $(Z_n)_{n\geq 0}$ *of real random variables on* Ω *given by*

$$Z_0 \overset{def}{=} 0, \quad Z_n \overset{def}{=} \sum_{k=1}^{n} X_k, \quad \forall n \geq 1.$$

Given any $n \in \mathbb{N}$ write $\sigma(X_1, \ldots, X_n)$ [resp. $\sigma(Z_1, \ldots, Z_n)$] for the σ-algebra generated by the first n variables of the sequence $(X_n)_{n\geq 1}$ [resp. the random walk $(Z_n)_{n\geq 1}$ generated by $(X_n)_{n\geq 1}$].

Proposition 992 *We have*

$$\sigma(Z_1, \ldots, Z_n) = \sigma(X_1, \ldots, X_n), \quad \forall n \in \mathbb{N}. \tag{11.33}$$

Proof. Since

$$Z_k = \sum_{j=1}^{k} X_j,$$

for every $k = 1, \ldots, n$, on account of Proposition 1423, we have

$$\sigma(Z_k) \subseteq \sigma(X_1, \ldots, X_n)$$

for every $k = 1, \ldots, n$. It Then, follows

$$\sigma(Z_1, \ldots, Z_n) = \bigvee_{k=1}^{n} \sigma(Z_k) \subseteq \sigma(X_1, \ldots, X_n).$$

Conversely, since

$$X_k = Z_k - Z_{k-1},$$

for every $k = 1, \ldots, n$, where $Z_0 = 0$, still on account of Proposition 1423, we have

$$\sigma(X_k) = \sigma(Z_0, Z_1, \ldots, Z_n) = \sigma(Z_1, \ldots, Z_n)$$

for every $k = 1, \ldots, n$. This implies

$$\sigma(X_1, \ldots, X_n) = \bigvee_{k=1}^{n} \sigma(X_k) \subseteq \sigma(Z_1, \ldots, Z_n)$$

and the desired (11.33) is completely proved. □

Proposition 993 *Assume $(X_n)_{n\geq 1}$ is a second-order sequence of uncorrelated random variables. Write $\mathbf{E}[X_n] \equiv \mu_n$ and $\mathbf{D}^2[X_n] \equiv \sigma_n^2$, where $\mu_n \in \mathbb{R}$ and $\sigma_n > 0$ for every $n \geq 1$, and consider the random walk $(Z_n)_{n\geq 0}$ generated by $(X_n)_{n\geq 1}$. We have*

$$\mathbf{E}[Z_n] = \sum_{k=1}^{n} \mu_k \quad \textit{and} \quad \mathbf{D}^2[Z_n] = \sum_{k=1}^{n} \sigma_k^2, \quad \forall n \geq 1.$$

Proof. The claim clearly follows form the linearity property of the expectation operator and from Proposition 989. □

Lemma 994 (Kolmogorov Inequality) *Assume $(X_n)_{n\geq 1}$ is a second-order sequence of independent random variables. Write $\mathbf{D}^2[X_n] \equiv \sigma_n^2$, where $\sigma_n > 0$ for every $n \geq 1$, and consider the random walk $(Z_n)_{n\geq 0}$ generated by $(X_n)_{n\geq 1}$. We have*

$$\mathbf{P}\left(\max_{1\leq k\leq n} |Z_k - \mathbf{E}[Z_k]| \geq \varepsilon\right) \leq \frac{1}{\varepsilon^2}\sum_{k=1}^{n} \sigma_k^2, \quad \forall \varepsilon > 0, \ \forall n \geq 1. \tag{11.34}$$

Proof. Note that for $n = 1$ the Kolomogorv inequality is just the Chebyshev inequality. Hence, assume $n > 1$ and consider the events

$$\begin{aligned}
E_1 &\equiv \{|Z_1 - \mathbf{E}[Z_1]| \geq \varepsilon\}, \\
E_2 &\equiv \{|Z_1 - \mathbf{E}[Z_1]| < \varepsilon \ \wedge \ |Z_2 - \mathbf{E}[Z_2]| \geq \varepsilon\}, \\
&\vdots \\
E_k &\equiv \{|Z_j - \mathbf{E}[Z_j]| < \varepsilon, \quad \forall j = 1, \ldots, k-1 \ \wedge \ |Z_k - \mathbf{E}[Z_k]| \geq \varepsilon\}, \\
&\vdots \\
E_n &\equiv \{|Z_j - \mathbf{E}[Z_j]| < \varepsilon, \quad \forall j = 1, \ldots, n-1 \ \wedge \ |Z_n - \mathbf{E}[Z_n]| \geq \varepsilon\},
\end{aligned}$$

for any fixed $\varepsilon > 0$. Note that, for any $k \leq n$, the event E_k is the set of the sample points $\omega \in \Omega$ such that $|Z_k - \mathbf{E}[Z_k]| \geq \varepsilon$ for the first time at the kth step of the random walk $(Z_n)_{n\geq 0}$. It is clearly seen that the events $E_1, \ldots, E_n$ are pairwise incompatible and we have

$$\left\{\max_{1\leq k\leq n} |Z_k - \mathbf{E}[Z_k]| \geq \varepsilon\right\} = \textstyle\bigcup_{k=1}^{n} E_k.$$

In addition, by virtue of (11.33) of Proposition 992, we have

$$E_k \in \sigma(Z_1, Z_2, \ldots, Z_k) = \sigma(X_1, X_2, \ldots, X_k),$$

for every $k = 1, \ldots, n$. Write $\mu_n \equiv \mathbf{E}[X_n]$, for every $n \geq 1$, and consider the random variables

$$X_n^* \stackrel{\text{def}}{=} X_n - \mu_n \quad \text{and} \quad Z_n^* \stackrel{\text{def}}{=} \sum_{k=1}^{n} X_k^*, \quad \forall n \geq 1.$$

We clearly have

$$\mathbf{E}[X_n^*] = 0, \qquad \mathbf{D}^2[X_n^*] = \mathbf{D}^2[X_n] = \sigma_n^2,$$

and

$$Z_n^* = Z_n - \mathbf{E}[Z_n], \qquad \mathbf{E}[Z_n^*] = 0,$$

for every $n \geq 1$. In addition,

$$\mathbf{D}^2 [Z_n^*] = \mathbf{D}^2 \left[\sum_{k=1}^n X_k^* \right] = \sum_{k=1}^n \mathbf{D}^2 [X_k^*] = \sum_{k=1}^n \sigma_k^2, \tag{11.35}$$

for every $n \geq 1$, because, under the assumption of independence of the random variables of the sequence $(X_n)_{n\geq 1}$, the random variables of the sequence $(X_n^*)_{n\geq 1}$ are independent as well. On the other hand, setting $E \equiv \{\max_{1\leq k\leq n} |Z_k - \mathbf{E}[Z_k]| \geq \varepsilon\}$, we have

$$\mathbf{D}^2 [Z_n^*] = \int_\Omega (Z_n^*)^2 \, d\mathbf{P} \geq \int_E (Z_n^*)^2 \, d\mathbf{P} = \int_{\bigcup_{k=1}^n E_k} (Z_n^*)^2 \, d\mathbf{P} = \sum_{k=1}^n \int_{E_k} (Z_n^*)^2 \, d\mathbf{P}, \tag{11.36}$$

and, thanks to the independence of $1_{E_k} Z_k^*$ and $\sum_{\ell=k+1}^n X_\ell^*$, we can write

$$\begin{aligned}
\int_{E_k} (Z_n^*)^2 \, d\mathbf{P} &= \int_{E_k} (Z_k^*)^2 + \left((Z_n^*)^2 - (Z_k^*)^2 \right) d\mathbf{P} \\
&= \int_{E_k} (Z_k^*)^2 \, d\mathbf{P} + \int_{E_k} \left(\left(Z_k^* + \sum_{\ell=k+1}^n X_\ell^* \right)^2 - (Z_k^*)^2 \right) d\mathbf{P} \\
&= \int_{E_k} (Z_k - \mathbf{E}[Z_k])^2 \, d\mathbf{P} + \int_{E_k} \left((Z_k^*)^2 + \left(\sum_{\ell=k+1}^n X_\ell^* \right)^2 + 2 Z_k^* \left(\sum_{\ell=k+1}^n X_\ell^* \right) - (Z_k^*)^2 \right) d\mathbf{P} \\
&\geq \int_{E_k} \varepsilon^2 d\mathbf{P} + 2 \int_{E_k} Z_k^* \left(\sum_{\ell=k+1}^n X_\ell^* \right) d\mathbf{P} + \int_{E_k} \left(\sum_{\ell=k+1}^n X_\ell^* \right)^2 d\mathbf{P} \\
&\geq \varepsilon^2 \mathbf{P}(E_k) + 2 \int_\Omega 1_{E_k} Z_k^* \left(\sum_{\ell=k+1}^n X_\ell^* \right) d\mathbf{P} \\
&= \varepsilon^2 \mathbf{P}(E_k) + 2\mathbf{E} \left[1_{E_k} Z_k^* \left(\sum_{\ell=k+1}^n X_\ell^* \right) \right] \\
&= \varepsilon^2 \mathbf{P}(E_k) + 2\mathbf{E}[1_{E_k} Z_k^*] \, \mathbf{E} \left[\sum_{\ell=k+1}^n X_\ell^* \right] \\
&= \varepsilon^2 \mathbf{P}(E_k).
\end{aligned} \tag{11.37}$$

Combining (11.35) and (11.37) it Then, follows,

$$\sum_{k=1}^n \sigma_k^2 \geq \varepsilon^2 \sum_{k=1}^n \mathbf{P}(E_k) = \varepsilon^2 \mathbf{P}(E),$$

which is the desired (11.34). □

Lemma 995 (Bernstein-Fréchet Inequality) *Assume $(X_n)_{n\geq 1}$ is a second-order sequence of independent random variables.*

1. *If there exist $a_k, b_k \in \mathbb{R}$ such that*

$$a_k \leq X_k \leq b_k, \quad \forall k = 1, \ldots, n,$$

we have

$$\mathbf{P}\left(\left|\sum_{k=1}^{n}\left(X_k-\mathbf{E}\left[X_k\right]\right)\right|\geq\varepsilon\right)\leq 2\exp\left(-\frac{2\varepsilon^2}{\sum_{k=1}^{n}\left(b_k-a_k\right)^2}\right),\quad\forall\varepsilon>0,\ \forall n\geq 1. \tag{11.38}$$

2. If there exist $C\in\mathbb{R}_+$ such that

$$\mathbf{E}\left(\left|X_k-\mathbf{E}\left[X_k\right]\right|^p\right)\leq C^{p-2}p!\mathbf{D}^2\left[X_k\right],\quad\forall p\geq 3,\ \forall k=1,\ldots,n,$$

we have

$$\mathbf{P}\left(\left|\sum_{k=1}^{n}\left(X_k-\mathbf{E}\left[X_k\right]\right)\right|\geq\varepsilon\right)\leq 2\exp\left(-\frac{\varepsilon^2}{4\sum_{k=1}^{n}\mathbf{D}^2\left[X_k\right]+2C\varepsilon}\right),\quad\forall\varepsilon>0,\ \forall n\geq 1. \tag{11.39}$$

Proof. . □

Let $(X_n)_{n\geq 1}$ be a sequence of real random variables on Ω.

Definition 996 *We say that $(X_n)_{n\geq 1}$ is a* sequence of identically distributed *random variables if all the random variables of the sequence have the same distribution. In symbols,*

$$P_{X_m}=P_{X_n},\quad\forall m,n\geq 1.$$

Clearly

Remark 997 *The random variables of the sequence $(X_n)_{n\geq 1}$ have the same distribution if and only if there exists a real random variable X on Ω such that*

$$P_{X_n}=P_X,\quad\forall n\geq 1.$$

Let $(Z_n)_{n\geq 0}$ be the *random walk* generated by $(X_n)_{n\geq 1}$.

Definition 998 *We call the* sample mean *of $(X_n)_{n\geq 1}$ the sequence $\left(\bar{X}_n\right)_{n\geq 1}$ of real random variables on Ω given by*

$$\bar{X}_n\overset{def}{=}\frac{Z_n}{n},\quad\forall n\geq 1.$$

Let $(X_n)_{n\geq 1}$ be a first-order sequence of real random variables on Ω.

Proposition 999 *Write $\mathbf{E}\left[X_n\right]\equiv\mu_n$, where $\mu_n\in\mathbb{R}$, for every $n\geq 1$. We have*

$$\mathbf{E}\left[\bar{X}_n\right]=\frac{1}{n}\sum_{k=1}^{n}\mu_n,\quad\forall n\geq 1. \tag{11.40}$$

Proof. In fact,

$$\mathbf{E}\left[\bar{X}_n\right]=\mathbf{E}\left[\frac{1}{n}\sum_{k=1}^{n}X_k\right]=\frac{1}{n}\sum_{k=1}^{n}\mathbf{E}\left[X_k\right]=\frac{1}{n}\sum_{k=1}^{n}\mu_n,$$

for every $n\geq 1$. □

Corollary 1000 *Assume* $(X_n)_{n\geq 1}$ *is a sequence of identically distributed random variables and write* $\mathbf{E}[X_n] \equiv \mu$*, where* $\mu \in \mathbb{R}$*, for every* $n \geq 1$*. Then*

$$\mathbf{E}\left[\bar{X}_n\right] = \mu. \tag{11.41}$$

Let $(X_n)_{n\geq 1}$ be a second-order sequence of real random variables on Ω.

Proposition 1001 *Assume the random variables of the sequence* $(X_n)_{n\geq 1}$ *have the same first-order moment and write* $\mathbf{E}[X_n] \equiv \mu$ *and* $\mathbf{D}^2[X_n] \equiv \sigma_n^2$*, where* $\mu \in \mathbb{R}$ *and* $\sigma_n > 0$ *for every* $n \geq 1$*. In addition, assume the random variables of the sequence* $(X_n)_{n\geq 1}$ *are uncorrelated. Then, we have*

$$\mathbf{D}^2\left[\bar{X}_n\right] = \frac{1}{n^2}\sum_{k=1}^{n}\sigma_k^2, \quad \forall n \geq 1. \tag{11.42}$$

Proof. In fact,

$$\mathbf{D}^2\left[\bar{X}_n\right] = \mathbf{D}^2\left[\frac{1}{n}Z_n\right] = \mathbf{D}^2\left[\frac{1}{n}\sum_{k=1}^{n}X_k\right] = \frac{1}{n^2}\mathbf{D}^2\left[\sum_{k=1}^{n}X_k\right] = \frac{1}{n^2}\sum_{k=1}^{n}\mathbf{D}^2[X_k] = \frac{1}{n^2}\sum_{k=1}^{n}\sigma_k^2,$$

for every $n \geq 1$. □

Corollary 1002 *Assume* $(X_n)_{n\geq 1}$ *is a sequence of identically distributed random variables and write* $\mathbf{E}[X_n] \equiv \mu$ *and* $\mathbf{D}^2[X_n] \equiv \sigma^2$*, where* $\mu \in \mathbb{R}$ *and* $\sigma > 0$ *for every* $n \geq 1$*. In addition, assume the random variables of the sequence* $(X_n)_{n\geq 1}$ *are uncorrelated. Then, we have*

$$\mathbf{D}^2\left[\bar{X}_n\right] = \frac{1}{n}\sigma. \tag{11.43}$$

Definition 1003 *We call the* unbiased *[resp.* biased*]* sample variance *of* $(X_n)_{n\geq 1}$ *the sequence* $(S_n^2)_{n\geq 1}$ *[resp.* $\left(\tilde{S}_n^2\right)_{n\geq 1}$*] of real random variables on* Ω *given by*

$$S_n^2 \overset{def}{=} \frac{\sum_{k=1}^{n}\left(X_k - \bar{X}_n\right)^2}{n-1}, \quad \forall n \geq 1 \quad \textit{[resp. } \tilde{S}_n^2 \overset{def}{=} \frac{\sum_{k=1}^{n}\left(X_k - \bar{X}_n\right)^2}{n}, \quad \forall n \geq 2\textit{]}.$$

Proposition 1004 *Assume* $(X_n)_{n\geq 1}$ *is a sequence of identically distributed random variables. Write* $\mathbf{E}[X_n] \equiv \mu$ *and* $\mathbf{D}^2[X_n] \equiv \sigma^2$*, where* $\mu \in \mathbb{R}$ *and* $\sigma > 0$ *for every* $n \geq 1$*. In addition, assume the random variables of the sequence* $(X_n)_{n\geq 1}$ *are uncorrelated. Then, we have*

$$\mathbf{E}\left[S_n^2\right] = \sigma^2 \quad \textit{and} \quad \mathbf{E}\left[\tilde{S}_n^2\right] = \frac{n-1}{n}\sigma^2, \quad \forall n \geq 1.$$

Proof. We have

$$\mathbf{E}\left[X_n^2\right] = \mathbf{D}^2[X_n] + \mathbf{E}[X_n]^2 = \sigma^2 + \mu^2,$$

and

$$\begin{aligned}
\mathbf{E}\left[\bar{X}_n^2\right] &= \mathbf{E}\left[\frac{1}{n^2}\left(\sum_{k=1}^{n} X_k\right)^2\right] = \frac{1}{n^2}\mathbf{E}\left[\sum_{k=1}^{n} X_k^2 + \sum_{\substack{j,k=1\\ j\neq k}}^{n} X_j X_k\right] \\
&= \frac{1}{n^2}\left(\sum_{k=1}^{n} \mathbf{E}\left[X_k^2\right] + \sum_{\substack{j,k=1\\ j\neq k}}^{n} \mathbf{E}\left[X_j X_k\right]\right) = \frac{1}{n^2}\left(\sum_{k=1}^{n} \mathbf{E}\left[X_k^2\right] + \sum_{\substack{j,k=1\\ j\neq k}}^{n} \mathbf{E}\left[X_j\right]\mathbf{E}\left[X_k\right]\right) \\
&= \frac{1}{n^2}\left(\sum_{k=1}^{n} \left(\sigma^2+\mu^2\right) + \sum_{\substack{j,k=1\\ j\neq k}}^{n} \mu^2\right) = \frac{1}{n^2}\left(n\left(\sigma^2+\mu^2\right) + n\left(n-1\right)\mu^2\right) \\
&= \frac{1}{n}\left(\sigma^2 + n\mu^2\right),
\end{aligned}$$

for every $n \geq 1$. In addition,

$$\begin{aligned}
\mathbf{E}\left[X_k \bar{X}_n\right] &= \mathbf{E}\left[X_k \frac{1}{n}\sum_{j=1}^{n} X_j\right] = \frac{1}{n}\sum_{j=1}^{n} \mathbf{E}\left[X_j X_k\right] = \frac{1}{n}\left(\mathbf{E}\left[X_k^2\right] + \sum_{\substack{j=1\\ j\neq k}}^{n} \mathbf{E}\left[X_j\right]\mathbf{E}\left[X_k\right]\right) \\
&= \frac{1}{n}\left(\sigma^2 + \mu^2 + \sum_{\substack{j=1\\ j\neq k}}^{n} \mu^2\right) = \frac{1}{n}\left(\sigma^2 + \mu^2 + \left(n-1\right)\mu^2\right) = \frac{1}{n}\left(\sigma^2 + n\mu^2\right),
\end{aligned}$$

for every $n \geq 1$ and for every $k = 1, \ldots, n$. As a consequence,

$$\begin{aligned}
\mathbf{E}\left[\tilde{S}_n^2\right] &= \mathbf{E}\left[\frac{\sum_{k=1}^{n}\left(X_k - \bar{X}_n\right)^2}{n}\right] = \frac{1}{n}\mathbf{E}\left[\sum_{k=1}^{n}\left(X_k - \bar{X}_n\right)^2\right] \\
&= \frac{1}{n}\mathbf{E}\left[\sum_{k=1}^{n}\left(X_k^2 + \bar{X}_n^2 - 2X_k\bar{X}_n\right)\right] \\
&= \frac{1}{n}\mathbf{E}\left[\sum_{k=1}^{n} X_k^2 + \sum_{k=1}^{n} \bar{X}_n^2 - 2\sum_{k=1}^{n} X_k\bar{X}_n\right] \\
&= \frac{1}{n}\left(\sum_{k=1}^{n} \mathbf{E}\left[X_k^2\right] + \sum_{k=1}^{n} \mathbf{E}\left[\bar{X}_n^2\right] - 2\sum_{k=1}^{n} \mathbf{E}\left[X_k\bar{X}_n\right]\right) \\
&= \frac{1}{n}\left(\sum_{k=1}^{n}\left(\sigma^2+\mu^2\right) + \sum_{k=1}^{n}\frac{1}{n}\left(\sigma^2+n\mu^2\right) - 2\sum_{k=1}^{n}\frac{1}{n}\left(\sigma^2+n\mu^2\right)\right) \\
&= \frac{1}{n}\left(n\left(\sigma^2+\mu^2\right) + \left(\sigma^2+n\mu^2\right) - 2\left(\sigma^2+n\mu^2\right)\right) \\
&= \frac{n-1}{n}\sigma^2,
\end{aligned}$$

for every $n \geq 1$. In the end, from the identity

$$\tilde{S}_n^2 = \frac{n-1}{n} S_n^2$$

which holds true for every $n \geq 2$, we have

$$\mathbf{E}\left[S_n^2\right] = \mathbf{E}\left[\frac{n}{n-1}\tilde{S}_n^2\right] = \frac{n}{n-1}\mathbf{E}\left[\tilde{S}_n^2\right] = \frac{n}{n-1}\frac{n-1}{n}\sigma^2 = \sigma^2.$$

This completes the proof of our claim. □

11.3 Weak Laws of Large Numbers

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space, let $(\mathbb{R}, \mathcal{B}(\mathbb{R})) \equiv \mathbb{R}$ be the Euclidean real line equipped with the Borel σ-algebra, let $(X_n)_{n\geq 1}$ be a sequence of real random variables on Ω, and let $\left(\bar{X}_n\right)_{n\geq 0}$ be the sample mean of $(X_n)_{n\geq 1}$.

Theorem 1005 (Bernoulli) *Assume $(X_n)_{n\geq 1}$ is a sequence of independent random variables which are Bernoulli distributed with success probability p. Then*

$$\bar{X}_n \xrightarrow{\boldsymbol{P}} p. \tag{11.44}$$

Proof. Note that we have $\mathbf{E}[X_n] = p$ and $\mathbf{D}^2[X_n] = pq$, for every $n \in \mathbb{N}$. It follows $\mathbf{E}\left[\bar{X}_n\right] = p$ and, thanks to the independence hypothesis on $(X_n)_{n\geq 1}$, we have

$$\mathbf{D}^2\left[\bar{X}_n\right] = \mathbf{D}^2\left[\frac{1}{n}\sum_{k=1}^{n} X_k\right] = \frac{1}{n^2}\sum_{k=1}^{n}\mathbf{D}^2[X_k] = \frac{1}{n^2}npq = \frac{pq}{n}.$$

Hence, by virtue of the Chebyshev inequality applied to $\bar{X}_n$ (see Corollary 656), it follows

$$\mathbf{P}\left(\left|\bar{X}_n - p\right| \geq \varepsilon\right) \leq \frac{\mathbf{D}^2\left[\bar{X}_n\right]}{\varepsilon^2} = \frac{pq}{n\varepsilon^2}.$$

which implies (11.44). □

Theorem 1006 *Assume $(X_n)_{n\geq 1}$ is a second-order sequence of independent random variables which have the same first-order moment. Write $\mathbf{E}[X_n] = \mu$ and $\mathbf{D}^2[X_n] = \sigma_n^2$, where $\mu \in \mathbb{R}$ and $\sigma_n > 0$ for every $n \geq 1$. In addition, assume*

$$\lim_{n\to\infty} \frac{1}{n^2}\textstyle\sum_{k=1}^{n}\sigma_k^2 = 0. \tag{11.45}$$

Then, we have

$$\bar{X}_n \xrightarrow{\boldsymbol{P}} \mu. \tag{11.46}$$

Proof. Considering (11.41) and (11.42), we have

$$\mathbf{E}\left[\bar{X}_n\right] = \mu \quad \text{and} \quad \mathbf{D}^2\left[\bar{X}_n\right] = \frac{1}{n^2}\textstyle\sum_{k=1}^{n}\sigma_k^2.$$

Therefore, thanks to the Chebyshev inequality applied to $\bar{X}_n$ (see Corollary 656), we can write

$$\mathbf{P}\left(\left|\bar{X}_n - \mu\right| \geq \varepsilon\right) \leq \frac{\sum_{k=1}^{n}\sigma_k^2}{n^2\varepsilon^2},$$

Hence, (11.46) clearly follows from (11.45). □

Corollary 1007 *Assume $(X_n)_{n\geq 1}$ is a second-order sequence of i.i.d. random variables. Then, we have*

$$\bar{X}_n \xrightarrow{\boldsymbol{P}} \mu. \tag{11.47}$$

Proof. We just observe that under the considered assumption (11.45) holds true. □

Theorem 1008 (Khintchine) *Assume $(X_n)_{n\geq 1}$ is a sequence of i.i.d. random variables of order 1. Write μ for the common value of $\mathbf{E}[X_n]$, on varying of $n \geq 1$. Then, we have*

$$\bar{X}_n \xrightarrow{\boldsymbol{P}} \mu.$$

Proof. . □

11.4 Strong Laws of Large Numbers

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space, let $(\mathbb{R}, \mathcal{B}(\mathbb{R})) \equiv \mathbb{R}$ be the Euclidean real line equipped with the Borel σ-algebra, let $(X_n)_{n\geq 1}$ be a sequence of real random variables on Ω, let $(Z_n)_{n\geq 0}$ be the random walk generated by $(X_n)_{n\geq 1}$, and let $\left(\bar{X}_n\right)_{n\geq 0}$ be the empirical mean of $(X_n)_{n\geq 1}$.

Theorem 1009 *Assume $(X_n)_{n\geq 1}$ is a second order sequence of independent random variables. Write $\mathbf{E}[X_n] \equiv \mu_n$ and $\mathbf{D}^2[X_n] \equiv \sigma_n^2$, where $\mu_n \in \mathbb{R}$ and $\sigma_n > 0$ for every $n \geq 1$. In addition, assume $\sum_{n=1}^{\infty} \sigma_n^2 < \infty$. Then, there exists $X \in L^2(\Omega; \mathbb{R})$ such that*

$$Z_n - \sum_{k=1}^{n} \mu_k \overset{a.s.}{\to} X.$$

Proof. ... . □

Theorem 1010 *Assume $(X_n)_{n\geq 1}$ is a second order sequence of independent random variables having the same first-order moment. Write $\mathbf{E}[X_n] = \mu$ and $\mathbf{D}^2[X_n] = \sigma_n^2$, where $\mu \in \mathbb{R}$ and $\sigma_n > 0$ for every $n \geq 1$. In addition, assume $\sum_{n=1}^{\infty} \frac{\sigma_n^2}{n^2} < \infty$. Then, we have*

$$\bar{X}_n \overset{a.s.}{\to} \mu.$$

Proof. ... . □

Corollary 1011 *Assume $(X_n)_{n\geq 1}$ is a second-order sequence of independent random variables having the same first and second order moment. Write $\mathbf{E}[X_n] = \mu$ and $\mathbf{D}^2[X_n] = \sigma^2$, where $\mu \in \mathbb{R}$ and $\sigma > 0$, for every $n \geq 1$. Then, we have*

$$\bar{X}_n \overset{a.s.}{\to} \mu.$$

Theorem 1012 (Kolmogorov) *Assume $(X_n)_{n\geq 1}$ is a first-order sequence of i.i.d. random variables. Write $\mathbf{E}[X_n] = \mu$. Then, we have*

$$\bar{X}_n \overset{a.s.}{\to} \mu.$$

Proof. For sake of simplicity assume $\mu = 0$. Set

$$Y_n \overset{\text{def}}{=} X_n 1_{\{|X_n| \leq n\}}, \quad Z_n \overset{\text{def}}{=} X_n 1_{\{|X_n| > n\}}, \quad \forall n \geq 1.$$

We have clearly,

$$X_n = Y_n + Z_n, \quad \forall n \geq 1.$$

Therefore, we will have proven the claim once we show that

$$\frac{1}{n}\sum_{k=1}^{n} Y_k \overset{\text{a.s.}}{\to} 0, \quad \frac{1}{n}\sum_{k=1}^{n} Z_k \overset{\text{a.s.}}{\to} 0.$$

By definition, we have

$$Z_n \neq 0 \Leftrightarrow |X_n| > n.$$

Furthermore, since $(X_n)_{n\geq 1}$ is a first order sequence of identically distributed random variables, there exists a real random variable $X \in L^1(\Omega; \mathbb{R})$ such that

$$P_{X_n} = P_X, \quad \forall n \geq 1.$$

First, consider $\sum_{n=1}^{\infty}\mathbf{P}\left(Z_n \neq 0\right)$. We have

$$\begin{aligned}\sum_{n=1}^{\infty}\mathbf{P}\left(Z_n \neq 0\right) &= \sum_{n=1}^{\infty}\mathbf{P}\left(|X_n| > n\right)\\ &= \sum_{n=1}^{\infty}\mathbf{P}\left(|X| > n\right)\\ &= \sum_{n=1}^{\infty}\mathbf{P}\left(\bigcup_{k=n}^{\infty}\left\{k < |X| \leq k+1\right\}\right)\\ &= \sum_{n=1}^{\infty}\sum_{k=n}^{\infty}\mathbf{P}\left(k < |X| \leq k+1\right)\\ &= \sum_{n=1}^{\infty} n\mathbf{P}\left(n < |X| \leq n+1\right).\end{aligned}$$

In fact, since the series $\sum_{n=1}^{\infty}\sum_{k=n}^{\infty}\mathbf{P}\left(k < |X| \leq k+1\right)$ has only positive terms, the reordering property applies and we can write

$$\begin{aligned}&\sum_{n=1}^{\infty}\sum_{k=n}^{\infty}\mathbf{P}\left(k < |X| \leq k+1\right)\\ &= \sum_{k=1}^{\infty}\mathbf{P}\left(k < |X| \leq k+1\right) + \sum_{k=2}^{\infty}\mathbf{P}\left(k < |X| \leq k+1\right) + \sum_{k=3}^{\infty}\mathbf{P}\left(k < |X| \leq k+1\right) + \ldots\\ &+ \sum_{k=n}^{\infty}\mathbf{P}\left(k < |X| \leq k+1\right) + \ldots\\ &= \mathbf{P}\left(1 < |X| \leq 2\right) + \mathbf{P}\left(2 < |X| \leq 3\right) + \mathbf{P}\left(3 < |X| \leq 4\right) + \cdots + \mathbf{P}\left(n < |X| \leq n+1\right) + \ldots\\ &+ \mathbf{P}\left(2 < |X| \leq 3\right) + \mathbf{P}\left(3 < |X| \leq 4\right) + \cdots + \mathbf{P}\left(n < |X| \leq n+1\right) + \ldots\\ &+ \mathbf{P}\left(3 < |X| \leq 4\right) + \cdots + \mathbf{P}\left(n < |X| \leq n+1\right) + \ldots\\ &+ \mathbf{P}\left(n < |X| \leq n+1\right) + \ldots\\ &= \mathbf{P}\left(1 < |X| \leq 2\right) + 2\mathbf{P}\left(2 < |X| \leq 3\right) + 3\mathbf{P}\left(3 < |X| \leq 4\right) + \cdots + n\mathbf{P}\left(n < |X| \leq n+1\right) + \ldots\\ &= \sum_{n=1}^{\infty} n\mathbf{P}\left(n < |X| \leq n+1\right).\end{aligned}$$

On the other hand,

$$\begin{aligned}\sum_{n=1}^{\infty} n\mathbf{P}\left(n < |X| \leq n+1\right) &= \sum_{n=1}^{\infty}\int_{\{n<|X|\leq n+1\}} n d\mathbf{P} \leq \sum_{n=1}^{\infty}\int_{\{n<|X|\leq n+1\}} |X|\, d\mathbf{P}\\ &\leq \int_{\Omega} |X|\, d\mathbf{P} = \mathbf{E}\left[|X|\right].\end{aligned}$$

Hence,

$$\sum_{n=1}^{\infty}\mathbf{P}\left(Z_n \neq 0\right) < \infty.$$

By virtue of the First Borel Cantelli Lemma (see **??** and **??**), we Then, have

$$\mathbf{P}\left(\limsup_{n\to\infty}\ \left\{Z_n \neq 0\right\}\right) = 0.$$

This means that the sample points $\omega \in \Omega$ which occur in infinitely many events of the sequence $\left(\left\{Z_n \neq 0\right\}\right)_{n\geq 1}$ constitute a neglegible event N. Hence, for every $\omega \in N^c$, which is an almost sure event, the set $\left\{Z_n\left(\omega\right) \neq 0\right\}$ is finite. As a consequence,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n} Z_k\left(\omega\right) = 0, \quad \forall \omega \in N^c.$$

Otherwise saying

$$\frac{1}{n}\sum_{k=1}^{n} Z_k \overset{\mathbf{a.s.}}{\to} 0.$$

Second, consider $(Y_n)_{n\geq 1}$. We have

$$\mathbf{E}[Y_n] = \mathbf{E}\left[X_n 1_{\{|X_n|\leq n\}}\right] = \int_\Omega X_n 1_{\{|X_n|\leq n\}} d\mathbf{P} = \int_{\mathbb{R}} x 1_{[-n,n]}(x)\, dP_{X_n}(x) = \int_{\mathbb{R}} x 1_{[-n,n]}(x)\, dP_X(x).$$

On the other hand,

$$\left|x 1_{[-n,n]}(x)\right| \leq |x|,$$

where x is Lebesgue integrable on $\mathbb{R}$ with respect to P_X because $X \in L^1(\Omega;\mathbb{R})$. Therefore, by virtue of the Lebesgue Dominated Convergence theorem, we have

$$\lim_{n\to\infty} \mathbf{E}[Y_n] = \lim_{n\to\infty} \int_{\mathbb{R}} x 1_{[-n,n]} dP_X = \int_{\mathbb{R}} \lim_{n\to\infty} x 1_{[-n,n]} dP_X = \int_{\mathbb{R}} x dP_X = \mathbf{E}[X] = 0.$$

As a consequence, thanks to the Cesàro Mean theorem, it follows

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^n \mathbf{E}[Y_k] = 0.$$

Moreover,

$$\begin{aligned}
\mathbf{E}\left[Y_n^2\right] &= \mathbf{E}\left[X_n^2 1_{\{|X_n|\leq n\}}\right] = \int_\Omega X_n^2 1_{\{|X_n|\leq n\}} d\mathbf{P} = \int_{\mathbb{R}} x^2 1_{[-n,n]}(x)\, dP_{X_n} \\
&= \int_{\mathbb{R}} x^2 1_{[-n,n]}(x)\, dP_X = \int_\Omega X^2 1_{\{|X|\leq n\}} d\mathbf{P} = \int_{\{|X|\leq n\}} X^2 d\mathbf{P} \\
&= \int_{\{|X|=0\}\cup\left(\bigcup_{k=0}^n \{k<|X|\leq k+1\}\right)} X^2 d\mathbf{P} = \int_{\{|X|=0\}} X^2 d\mathbf{P} + \int_{\bigcup_{k=0}^n \{k<|X|\leq k+1\}} X^2 d\mathbf{P} \\
&= \sum_{k=0}^{n-1} \int_{\{k<|X|\leq k+1\}} X^2 d\mathbf{P} \leq \sum_{k=0}^{n-1} (k+1)^2\, \mathbf{P}(k<|X|\leq k+1)
\end{aligned}$$

On the other hand, $\mathbf{D}^2[Y_n] \leq \mathbf{E}[Y_n^2]$. Hence, we can write

$$\sum_{n=1}^\infty \frac{\mathbf{D}^2[Y_n]}{n^2} \leq \sum_{n=1}^\infty \frac{\mathbf{E}[Y_n^2]}{n^2} \leq \sum_{n=1}^\infty \frac{1}{n^2} \left(\sum_{k=0}^{n-1} (k+1)^2\, \mathbf{P}(k<|X|\leq k+1)\right).$$

As above, since the series $\sum_{n=1}^{\infty}\frac{1}{n^2}\left(\sum_{k=0}^{n-1}(k+1)^2\mathbf{P}(k<|X|\leq k+1)\right)$ has only positive terms, the reordering property applies and we can write

$$\begin{aligned}
&\sum_{n=1}^{\infty}\frac{1}{n^2}\left(\sum_{k=0}^{n-1}(k+1)^2\mathbf{P}(k<|X|\leq k+1)\right)\\
&=\sum_{k=0}^{0}(k+1)^2\mathbf{P}(k<|X|\leq k+1)+\frac{1}{4}\sum_{k=0}^{1}(k+1)^2\mathbf{P}(k<|X|\leq k+1)\\
&+\frac{1}{9}\sum_{k=0}^{2}(k+1)^2\mathbf{P}(k<|X|\leq k+1)+\ldots\\
&+\frac{1}{n^2}\left(\sum_{k=0}^{n-1}(k+1)^2\mathbf{P}(k<|X|\leq k+1)\right)+\ldots\\
&=\mathbf{P}(0<|X|\leq 1)\\
&+\frac{1}{4}\left(\mathbf{P}(0<|X|\leq 1)+4\mathbf{P}(1<|X|\leq 2)\right)\\
&+\frac{1}{9}\left(\mathbf{P}(0<|X|\leq 1)+4\mathbf{P}(1<|X|\leq 2)+9\mathbf{P}(2<|X|\leq 3)\right)+\ldots\\
&+\frac{1}{n^2}\left(\mathbf{P}(0<|X|\leq 1)+4\mathbf{P}(1<|X|\leq 2)+9\mathbf{P}(2<|X|\leq 3)+\cdots+n^2\mathbf{P}(n-1<|X|\leq n)\right)+\ldots\\
&=\mathbf{P}(0<|X|\leq 1)\left(\sum_{m=1}^{\infty}\frac{1}{m^2}\right)+4\mathbf{P}(1<|X|\leq 2)\left(\sum_{m=2}^{\infty}\frac{1}{m^2}\right)\\
&+9\mathbf{P}(2<|X|\leq 3)\left(\sum_{m=3}^{\infty}\frac{1}{m^2}\right)+\ldots\\
&+n^2\mathbf{P}(n-1<|X|\leq n)\left(\sum_{m=n}^{\infty}\frac{1}{m^2}\right)+\ldots\\
&=\sum_{k=1}^{\infty}k^2\mathbf{P}(k-1<|X|\leq k)\left(\sum_{n=k}^{\infty}\frac{1}{n^2}\right).
\end{aligned}$$

Now, it is well known that,

$$\sum_{n=k}^{\infty}\frac{1}{n^2}\leq\frac{2}{k},\quad\forall n\geq 1.$$

Therefore,

$$\begin{aligned}
\sum_{n=1}^{\infty}\frac{\mathbf{D}^2[Y_n]}{n^2}&\leq\sum_{k=1}^{\infty}k^2\mathbf{P}(k-1<|X|\leq k)\frac{2}{k}\\
&=2\sum_{k=1}^{\infty}k\mathbf{P}(k-1<|X|\leq k)\\
&=2\sum_{k=1}^{\infty}(1+(k-1))\mathbf{P}(k-1<|X|\leq k)\\
&=2\left(\sum_{k=1}^{\infty}\mathbf{P}(k-1<|X|\leq k)+\sum_{k=1}^{\infty}(k-1)\mathbf{P}(k-1<|X|\leq k)\right)\\
&\leq 2\left(1+\sum_{k=1}^{\infty}\int_{\{k-1<|X|\leq k\}}(k-1)\,d\mathbf{P}\right)\\
&\leq 2\left(1+\int_{\Omega}|X|\,dP\right)\\
&=2(1+\mathbf{E}[X])
\end{aligned}$$

In the end, the Kolmogorov Lemma applies and we obtain

$$\frac{1}{n}\sum_{k=1}^{n}Y_k\overset{\mathbf{a.s.}}{\to}0.$$

as desired. □

Proposition 1013 *Assume $(X_n)_{n\geq 1}$ is a second-order sequence of i.i.d. random variables such that $\mathbf{E}[X_n] = 0$. Write $\mathbf{D}^2[X_n] \equiv \sigma^2$, where $\sigma > 0$ for every $n \geq 1$, and consider the empirical biased variance $\left(\tilde{S}_n^2\right)_{n\geq 1}$ of $(X_n)_{n\geq 1}$. Then, we have*

$$\tilde{S}_n^2 \overset{a.s.}{\to} \sigma^2.$$

Proof. ... . □

Theorem 1014 *Assume $(X_n)_{n\geq 1}$ is a second-order sequence of uncorrelated random variables such that $\mathbf{E}[X_n] = \mu$ and $\mathbf{D}^2[X_n] \leq \nu$, where $\mu \in \mathbb{R}$ and $\nu > 0$ for every $n \geq 1$. Then, we have*

$$\bar{X}_n \overset{a.s.}{\to} \mu.$$

Proof. Without loss in generality, we can suppose

$$\mathbf{E}[X_n] = 0, \quad \forall n \geq 1, \tag{11.48}$$

Hence, under our hypotheses, it is immediately seen that

$$\mathbf{D}^2[Z_n] = \mathbf{E}\left[Z_n^2\right] - \mathbf{E}[Z_n]^2 = \sum_{k=1}^n \mathbf{E}\left[X_k^2\right] = \sum_{k=1}^n \mathbf{D}^2[X_k] \leq n\nu, \quad \forall n \geq 1.$$

By the Chebyshev inequality, for every $\varepsilon > 0$ we have

$$\mathbf{P}(|Z_n| > \varepsilon n) \leq \frac{\mathbf{D}^2[Z_n]}{\varepsilon^2 n^2} = \frac{\nu}{\varepsilon^2 n}$$

for every $n \geq 1$. In particular,

$$\mathbf{P}\left(|Z_{n^2}| > \varepsilon n^2\right) \leq \frac{\nu}{\varepsilon^2 n^2},$$

for every $n \geq 1$. It Then, follows that the series

$$\sum_{n=1}^{\infty} \mathbf{P}\left(\frac{|Z_{n^2}|}{n^2} > \varepsilon\right)$$

converges, and, by virtue of Corollary 927, we obtain

$$\bar{X}_{n^2} = \frac{Z_{n^2}}{n^2} \overset{\mathbf{a.s.}}{\to} 0. \tag{11.49}$$

Thus, we have proved the claim for a subsequence of $\left(\bar{X}_n\right)_{n\geq 1}$. We will obtain the desired result for the whole sequence by showing that elements of $(Z_n)_{n\geq 1}$ do not differ enough from the nearest Z_{n^2} to make any real difference. To this, write

$$D_n \overset{\text{def}}{=} \max_{n^2+1\leq k\leq (n+1)^2} |Z_k - Z_{n^2}|, \quad \forall n \geq 1.$$

On account of (11.48), the uncorrelation assumption on $(X_n)_{n\geq 1}$ implies that for every $n^2 + 1 \leq k \leq (n+1)^2$

$$\mathbf{E}\left[(Z_k - Z_{n^2})^2\right] = \mathbf{E}\left[\left(\sum_{j=n^2+1}^{k} X_j\right)^2\right] = \sum_{j=n^2+1}^{k} \mathbf{E}\left[X_j^2\right] \leq \sum_{j=n^2+1}^{(n+1)^2} \mathbf{E}\left[X_j^2\right] \leq (2n+1)\, M.$$

Hence,

$$\mathbf{E}\left[D_n^2\right] \leq (2n+1)^2 M \leq 9n^2 M.$$

By virtue of the Chebyshev's inequality, we Then, obtain

$$\mathbf{P}\left(|D_n| > \varepsilon n^2\right) \leq \frac{9M}{\varepsilon^2 n^2},$$

and it follows as before

$$\frac{D_n}{n^2} \overset{\text{a.s.}}{\to} 0. \tag{11.50}$$

Now, since for every $n^2 + 1 \leq k \leq (n+1)^2$ we have

$$\left|\bar{X}_n\right| = \frac{|Z_k|}{k} \leq \frac{|Z_{n^2}| + D_n}{n^2}$$

it is clear that (11.49) and (11.50) together yield the desired result. □

Theorem 1015 (Kolmogorov three series theorem) *Assume $(X_n)_{n\geq 1}$ is a second-order sequence of uncorrelated random variables. Then, there exists $X \in L^2(\Omega;\mathbb{R})$ such that*

$$Z_n \overset{\text{a.s.}}{\to} X.$$

if and only if

$$\sum_{n=1}^{\infty}\mathbf{P}\left(|X_n| > 1\right) < \infty, \qquad \sum_{n=1}^{\infty}\mathbf{E}\left[X_n 1_{\{|X_n|\leq 1\}}\right] < \infty, \qquad \sum_{n=1}^{\infty}\mathbf{D}^2\left[X_n 1_{\{|X_n|\leq 1\}}\right] < \infty.$$

Proof. . □

11.5 Laws of Large Numbers in L^p

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space, let $(\mathbb{R}, \mathcal{B}(\mathbb{R})) \equiv \mathbb{R}$ be the Euclidean real line equipped with the Borel σ-algebra, let $(X_n)_{n\geq 1}$ be a sequence of real random variables on Ω, let $(Z_n)_{n\geq 0}$ be the random walk generated by $(X_n)_{n\geq 1}$, and let $\left(\bar{X}_n\right)_{n\geq 0}$ be the empirical mean of $(X_n)_{n\geq 1}$.

Theorem 1016 *Assume $(X_n)_{n\geq 1}$ is a second-order sequence of uncorrelated random variables such that $\mathbf{E}[X_n] = \mu$ and $\mathbf{D}^2[X_n] \leq \nu$, where $\mu \in \mathbb{R}$ and $\nu > 0$ for every $n \geq 1$. We Then, have*

$$\bar{X}_n \overset{L^2}{\to} \mu.$$

Proof. The uncorrelation assumption allows us to write

$$\begin{aligned}\left\|\bar{X}_n - \mu\right\|_2 &= \left\|\frac{Z_n - \mathbf{E}[Z_n]}{n}\right\|_2 = \frac{1}{n}\left\|Z_n - \mathbf{E}[Z_n]\right\|_2 = \frac{1}{n}\left(\int_\Omega \left(Z_n - \mathbf{E}[Z_n]\right)^2 d\mathbf{P}\right)^{1/2} \\ &= \frac{1}{n}\left(\mathbf{D}^2[Z_n]\right)^{1/2} = \frac{1}{n}\left(\sum_{k=1}^{n}\mathbf{D}^2[X_k]\right)^{1/2} \leq \frac{1}{n}(n\nu)^{1/2} = \frac{\nu^{1/2}}{n^{1/2}},\end{aligned}$$

which yields the the desired result. □

Theorem 1017 *Assume $(X_n)_{n\geq 1}$ is a second-order sequence of uncorrelated random variables such that $\mathbf{E}[X_n] = 0$. Write $\mathbf{D}^2[X_n] = \mathbf{E}^2[X_n] \equiv \sigma_n^2$, where $\sigma_n > 0$ for every $n \geq 1$. In addition, assume that $\sum_{n=1}^{\infty}\sigma_n^2 < \infty$. Then, there exists $X \in L^2(\Omega;\mathbb{R})$ such that we have*

$$Z_n \xrightarrow{\boldsymbol{L}^2} X$$

and

$$\mathbf{E}\left[\left(\sum_{n=1}^{\infty}X_n^2\right)\right] = \sum_{n=1}^{\infty}\mathbf{E}\left[X_n^2\right].$$

Proof. ... . □

Theorem 1018 *Assume $(X_n)_{n\geq 1}$ is a first-order sequence of i.i.d. random variables. We Then, have*

$$\left|\bar{X}_n\right| \xrightarrow{\boldsymbol{L}^1} |\mathbf{E}[X]|.$$

Proof. ... . □

11.6 Central Limit Theorem

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space and let $(\mathbb{R}, \mathcal{B}(\mathbb{R})) \equiv \mathbb{R}$ be the Euclidean real line equipped with the Borel σ-algebra. Let $(X_n)_{n\geq 1}$ be a sequence of real random variables on Ω and let $(Z_n)_{n\geq 0}$ be the random walk generated by $(X_n)_{n\geq 1}$.

Theorem 1019 (central limit theorem) *Assume $(X_n)_{n\geq 1}$ is a second-order sequence of i.i.d. random variables. Write $\mathbf{E}[X_n] = \mu$ and $\mathbf{D}^2[X_n] = \sigma^2$, where $\mu \in \mathbb{R}$ and $\sigma > 0$, and consider the standardized random walk $\left(\tilde{Z}_n\right)_{n\geq 1}$ given by*

$$\tilde{Z}_n \overset{def}{=} \frac{Z_n - n\mu}{\sigma\sqrt{n}}, \quad \forall n \geq 1.$$

Then, $\left(\tilde{Z}_n\right)_{n\geq 1}$ converges in distribution to a standard normal random variable. In symbols

$$\tilde{Z}_n \xrightarrow{w} N(0,1).$$

Proof. Write

$$\tilde{X}_n \equiv \frac{X_n - \mu}{\sigma}, \quad n \geq 1.$$

It is clearly seen that the random variables of the sequence $\left(\tilde{X}_n\right)_{n\geq 1}$ fulfill the same assumptions on the random variables of the sequence $(X_n)_{n\geq 1}$. Moreover,

$$\mathbf{E}\left[\tilde{X}_n\right] = 0, \quad \mathbf{E}\left[\tilde{X}_n^2\right] = \mathbf{D}^2\left[\tilde{X}_n\right] = 1, \quad \text{and} \quad \tilde{Z}_n = \frac{1}{\sqrt{n}}\sum_{k=1}^n \tilde{X}_k,$$

for every $n \geq 1$. Now, consider the characteristic function of $\tilde{Z}_n$. By virtue of Propositions 814 and 825, we have

$$\varphi_{\tilde{Z}_n}(t) = \prod_{k=1}^n \varphi_{\tilde{X}_k}\left(\frac{t}{\sqrt{n}}\right) = \varphi_{\tilde{X}}\left(\frac{t}{\sqrt{n}}\right)^n$$

where $\tilde{X}$ is any random variable of the sequence $\left(\tilde{X}_n\right)_{n\geq 1}$. On the other hand, thanks to Theorem 819 the function $\varphi_{\tilde{X}}$ is two times differentiable and we can write

$$\begin{aligned}
\varphi_{\tilde{X}}\left(\frac{t}{\sqrt{n}}\right) &= \varphi_{\tilde{X}}(0) + \frac{1}{1!}\varphi'_{\tilde{X}}(0)\left(\frac{t}{\sqrt{n}}\right) + \frac{1}{2!}\varphi''_{\tilde{X}}(0)\left(\frac{t}{\sqrt{n}}\right)^2 + o\left(\left(\frac{t}{\sqrt{n}}\right)^2\right) \\
&= 1 + i\mathbf{E}\left[\tilde{X}\right]\left(\frac{t}{\sqrt{n}}\right) + i^2\mathbf{E}\left[\tilde{X}^2\right]\left(\frac{t}{\sqrt{n}}\right)^2 + o\left(\frac{t^2}{n}\right) \\
&= 1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right).
\end{aligned}$$

Hence,

$$\varphi_{\tilde{Z}_n}(t) = \left(1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n,$$

Thus, on account that

$$\lim_{n\to\infty} o\left(\frac{t^2}{n}\right) = 0, \quad \forall t \in \mathbb{R},$$

we have

$$\lim_{n\to\infty}\varphi_{\tilde{Z}_n}(t)=\lim_{n\to\infty}\left(1-\frac{t^2}{2n}+o\left(\frac{t^2}{n}\right)\right)^n=e^{-\frac{t^2}{2}}.$$

In the end, since $e^{-\frac{t^2}{2}}$ is the characteristic function of an $N(0,1)$ random variable, by virtue of Corollary 953 to the Levy Theorem 952, we obtain the desired result. □

Theorem 1020 *Assume $(X_n)_{n\geq 1}$ is a sequence of independent random variables such that $X_n \sim Bin\left(n,\frac{p}{n}\right)$. Then, the sequence $(X_n)_{n\geq 1}$ converges in distribution to a Poisson random variable of parameter p. In symbols*

$$X_n \xrightarrow{w} Poiss(p).$$

Proof. ... . □

11.7 Kolmogorov 0-1 Law

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space, let $\left(\bar{\mathbb{R}}, \mathcal{B}\left(\bar{\mathbb{R}}\right)\right) \equiv \bar{\mathbb{R}}$ be the extended Euclidean real line equipped with the Borel σ-algebra and let $(X_n)_{n\geq 1}$ be a sequence of extended real random variables on Ω.

Definition 1021 *Write $\mathcal{T}_n \equiv \sigma(X_{n+1}, X_{n+2}, \dots)$ for every $n \geq 1$. We call the* tail σ-algebra *generated by $(X_n)_{n\geq 1}$ the family $\mathcal{T}$ of events given by*

$$\mathcal{T} \overset{def}{=} \bigcap_{n\geq 1} \mathcal{T}_n.$$

Proposition 1022 *The maps $\limsup_{n\to\infty} X_n : \Omega \to \bar{\mathbb{R}}$ and $\liminf_{n\to\infty} X_n : \Omega \to \bar{\mathbb{R}}$ are $\left(\mathcal{T}, \mathcal{B}\left(\bar{\mathbb{R}}\right)\right)$-random variables.*

Proof. Consider the families of left-bounded and right-bounded extended-real half lines

$$\mathcal{H}_{l.c.\mathbb{Q}}\left(\bar{\mathbb{R}}\right) \equiv \{[a, +\infty], \quad a \in \mathbb{Q}\} \quad \text{and} \quad \mathcal{H}_{r.c.\mathbb{Q}}\left(\bar{\mathbb{R}}\right) \equiv \{[-\infty, a], \quad a \in \mathbb{Q}\},$$

respectively, which are well known to be a basis for the Borel σ-algebra $\mathcal{B}\left(\bar{\mathbb{R}}\right)$. Setting

$$\overline{L}_m \equiv \sup_{n\geq m} X_n, \quad \forall m \geq 1, \quad \text{and} \quad \underline{L}_m \equiv \inf_{n\geq m} X_n, \quad \forall m \geq 1,$$

we have

$$\limsup_{n\to\infty} X_n = \inf_{m\geq 1} \overline{L}_m \equiv \overline{L} \quad \text{and} \quad \liminf_{n\to\infty} X_n = \sup_{m\geq 1} \underline{L}_m \equiv \underline{L}$$

Now,

$$\overline{L}_m \in [a, +\infty] = \bigcap_{k\geq 1} \bigcup_{n\geq m} \left\{X_n \in (a - \tfrac{1}{k}, +\infty]\right\}, \quad \forall a \in \mathbb{Q},$$

and

$$\underline{L}_m \in [-\infty, a] = \bigcap_{k\geq 1} \bigcup_{n\geq m} \left\{X_n \in [-\infty, a + \tfrac{1}{k})\right\}, \quad \forall a \in \mathbb{Q}.$$

Hence, $\overline{L}_m \in [a, +\infty]$ and $\underline{L}_m \in [-\infty, a]$ are in $\mathcal{T}_{m-1}$ for every $m \geq 1$, since $\bigcup_{n\geq m} \left\{X_n \in (a - \frac{1}{k}, +\infty]\right\}$ and $\bigcup_{n\geq m} \left\{X_n \in [-\infty, a + \frac{1}{k})\right\}$ do, for every $k \geq 1$. It Then, follows that

$$\left\{\overline{L} \in [a, +\infty]\right\} = \left\{\inf_{m\geq 1} \overline{L}_m \in [a, +\infty]\right\} = \bigcap_{m\geq 1} \left\{\overline{L}_m \in [a, +\infty]\right\}$$

and

$$\left\{\underline{L} \in [-\infty, a]\right\} = \left\{\sup_{m\geq 1} \underline{L}_m \in [-\infty, a]\right\} = \bigcap_{m\geq 1} \left\{\underline{L}_m \in [-\infty, a]\right\}$$

are in $\mathcal{T}$. □

Let $(X_n)_{n\geq 1}$ is a sequence of independent extended real random variables and let $\mathcal{T}$ be the *tail σ-algebra* generated by $(X_n)_{n\geq 1}$.

Theorem 1023 (Kolmogorov 0-1 Law) *We have*

$$\mathbf{P}(E) = 0 \quad or \quad \mathbf{P}(E) = 1, \quad \forall E \in \mathcal{T}.$$

Moreover, every $\left(\mathcal{T}, \mathcal{B}\left(\bar{\mathbb{R}}\right)\right)$*-random variable is deterministic.*

Proof. Consider the σ-algebra of events

$$\mathcal{F}_n \equiv \sigma(X_1, \ldots, X_n), \quad \forall n \geq 1.$$

We have

$$\mathcal{F}_n = \sigma\left(\{E_{a_1,\ldots,a_n}, \quad \forall a_1, \ldots, a_n \in \mathbb{R}\}\right),$$

where

$$E_{a_1,\ldots,a_n} \equiv \{X_1 \leq a_1, \ldots, X_n \leq a_n\}.$$

On the other hand, we have

$$\mathcal{T}_n = \sigma\left(\left\{E_{a_{n+1},a_{n+2},\ldots}, \quad \forall a_{n+1}, a_{n+2}, \cdots \in \mathbb{R}\right\}\right),$$

where

$$E_{a_{n+1},a_{n+2},\ldots} \equiv \{X_{n+1} \leq a_{n+1}, X_{n+2} \leq a_{n+2}, \ldots\}.$$

Therefore, the independence of the random variables of the sequence $(X_n)_{n\geq 1}$ makes $\mathcal{F}_n$ and $\mathcal{T}_n$ independent, for every $n \geq 1$. It Then, follows that $\mathcal{T} \equiv \bigcap_{n\geq 1} \mathcal{T}_n$ is independent of $\mathcal{F}_n$, for every $n \geq 1$. On the other hand, $\mathcal{F}_\infty \equiv \sigma\left(\bigcup_{n\geq 1} \mathcal{F}_n\right)$ is generated by events in the families $\mathcal{F}_n$. Therefore, $\mathcal{T}$ is independent also of $\mathcal{F}_\infty$. Nevertheless, we have

$$\mathcal{T} \subseteq \mathcal{F}_\infty.$$

Hence, any event $E \in \mathcal{T}$ is independent of itself. We Then, have

$$\mathbf{P}(E) = \mathbf{P}(E \cap E) = \mathbf{P}(E)\,\mathbf{P}(E),$$

which implies the desired result. □

Part IV

Elements of Statistics

Chapter 12

Populations, Samples, Statistics

In inferential statistics, by observing a characteristic of the elements of a subset of a given set, we aim to infer the same characteristic about the elements of the whole set. The set under concern is called a *population*. The selected subset is called a *sample* drawn from the population. The number of elements in the population [resp. sample] is called the *size* of the population [resp. sample]. We hardly ever are in a position to know the values taken by the characteristic of interest on all population elements. We often do not even know some summary measures. Therefore, it is natural to represent such a characteristic by a random variable whose distribution may be unknown or may present summary measures expressed by parameters whose true value is unknown. In this context, we call a *statistic* a suitable function of the selected sample which allows to infer the type of the distribution or estimate the true values of the parameters while assessing the precision of these estimates. For instance, we could be interested in studying the height of the European women whose age is between 20 and 70. In this case, the population consists obviously of all European women whose age is between 20 and 70. The trait under consideration is just the height of the individuals of the population. The challenge is to draw a small *representative* sample from the population such that the observation of the height on the individuals of this sample allows us to infer the distribution of the height on the entire population or to estimate the average height and compute the error that we may commit when replacing our estimate to the true value. Note that the term *representative* does not mean that the distribution of the trait on the elements of the sample is the same than the distribution on the elements of the population. Rather, it means that the sample is chosen in such a way that any elements of the population has an equal chance to be included in the sample. It is more likely to obtain a representative sample by choosing the elements of a sample of a population in a totally random fashion than by prior considerations. This is because any specific non-random rule for selecting the elements of a sample results in a sample that is inherently biased towards some data values. In other words, we should not try to select a sample in the attempt to build a miniature copy of the population. We should leave this selection to the chance.

Definition 1024 *A sample of n members selected from a population of N members is said to be a* random sample*, if the members of the sample are chosen in such a way that all possible choices among the N members of the population are equally likely. The number n [resp. N] of elements in the sample [resp. population] is called the* size *of the sample [resp. population].*

A random sample has to be selected in such a way that it could be any of the possible samples of size n from the population of size N. Once a random sample is selected, such a

sample is representative: we can use statistical inference to draw conlusions about the entire population by studying the elements of the sample.

Example 1025 *On September* 2016 *we knew that a referendum would have been held on the* 4*th of December* 2016*. Different groups of students of a Data Science Master class aimed to predict whether the Yes or No would have prevailed. To this goal, the students wondered how to select and poll a sample of the voting population. Which of the following methods would have resulted in a representative sample? Why?*

1. *Poll all people of voting age attenting on the* 18*th of September at the Fiorentina-Roma football match.*
2. *Poll all people of voting age enjoying shopping the* 18*th of September at the Tor Vergata mall.*
3. *Poll randomly choosen people from the telephone directory.*
4. *Poll the respondents of randomly dialed mobile phone numbers.*
5. *Obtain a copy of the national voter registration list, randomly choose* 20.000 *names and poll them.*
6. *Obtain a copy of each regional voter registration list, randomly choose* 1.000 *names from each of the* 20 *list and poll them.*
7. *Another idea?*

12.0.1 Random sampling models

(a) sampling from an infinite population: the samples are independent and identically distributed;

(b) sampling with replacement from a finite population: the samples are independent and identically distributed;

(c) sampling without replacement from a finite population: the samples are not independent but still identically distributed.

Proposition 1026 *Assume the size N of a finite population is large if compared to the size n of the sample. Then, in sampling without replacement the samples will be approximately independent.*

Proof. . □

Chapter 13

Statistics on Simple Random Samples

Let X be a real random variable on a probability space $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ representing a population.

Definition 1027 (simple random sample) *Given any $n \in \mathbb{N}$, we say that the random variables $X_1, \ldots, X_n$ constitute a* simple random sample of size n drawn from X*, if $X_1, \ldots, X_n$ are independent and each X_k has the same distribution of X. In symbols*

$$X_k \perp\!\!\!\perp X_\ell \quad \textit{and} \quad X_k \overset{d}{=} X,$$

for all $k, \ell = 1, \ldots, n$, $\ell \neq k$.

Let $n \in \mathbb{N}$ and let $X_1, \ldots, X_n$ be a simple random sample of size n drawn from X.

Definition 1028 (statistic on a simple random sample) *Given any Borel function $g : \mathbb{R}^n \to \mathbb{R}$, the real random variable $G_n : \Omega \to \mathbb{R}$, briefly G_n, given by*

$$G_n \overset{def}{=} g \circ (X_1, \ldots, X_n), \tag{13.1}$$

or, more explicitly, by

$$G_n(\omega) \overset{def}{=} g(X_1(\omega), \ldots, X_n(\omega)), \quad \forall \omega \in \Omega, \tag{13.2}$$

is said to be a statistic on $X_1, \ldots, X_n$ *or a* statistic of size n drawn from X.

Definition 1029 (sample sum) *Consider the Borel function $g : \mathbb{R}^n \to \mathbb{R}$, given by*

$$g(x_1, \ldots, x_n) \overset{def}{=} \sum_{k=1}^{n} x_k, \quad \forall (x_1, \ldots, x_n) \in \mathbb{R}^n. \tag{13.3}$$

We call the statistic

$$Z_n \overset{def}{=} g \circ (X_1, \ldots, X_n) \equiv \sum_{k=1}^{n} X_k, \tag{13.4}$$

the sample sum *of size n drawn from X.*

Proposition 1030 (mean of the sample sum) *Assume that X has finite moment of order 1. Write $\mathbf{E}[X] \equiv \mu_X$. Then, Z_n has finite moment of order 1. Moerover, we have*

$$\mathbf{E}[Z_n] \equiv \mu_{Z_n} = n\mu_X. \tag{13.5}$$

Proof. Since $X_k \overset{\mathrm{d}}{=} X$, the random variable X_k has also finite moment of order 1 and we have

$$\mathbf{E}[X_k] = \mathbf{E}[X] \equiv \mu_X,$$

for every $k = 1, \ldots, n$. Since the sum of any finite number of random variables with finite moment of order 1 has also finite momment of order 1, it follows that the random variable Z_n has finite moment of order 1. In the end, thanks to the linearity property of the expectation operator, we have

$$\mathbf{E}[Z_n] = \mathbf{E}\left[\sum_{k=1}^{n} X_k\right] = \sum_{k=1}^{n} \mathbf{E}[X_k] = \sum_{k=1}^{n} \mu_X = n\mu_X,$$

as desired. □

Proposition 1031 (variance of the sample sum) *Assume that X has finite moment of order 2. Write $\mathbf{D}^2[X] \equiv \sigma_X^2$. Then, Z_n has finite moment of order 2. Moreover, we have*

$$\mathbf{D}^2[Z_n] \equiv \sigma_{Z_n}^2 = n\sigma_X^2. \tag{13.6}$$

Proof. Mutatis mutandis in the first part of the proof of Proposition (1030), we obtain that X_k has finite moment of order 2 and we have

$$\mathbf{D}^2[X_k] = \mathbf{D}^2[X] \equiv \sigma_X,$$

for every $k = 1, \ldots, n$. Moreover, Z_n has finite moment of order 2. Hence, thanks to the additivity property of the variance operator on a sum of independent random variables, we can write

$$\mathbf{D}^2[Z_n] = \mathbf{D}^2\left[\sum_{k=1}^{n} X_k\right] = \sum_{k=1}^{n} \mathbf{D}^2[X_k] = \sum_{k=1}^{n} \sigma_X^2 = n\sigma_X^2.$$

This completes the proof. □

Remark 1032 *Assume that X is Bernoulli distributed with success probability p. Then, Z_n is binomially distributed with number of trials parameter n and success probabilitry p. In symbols,*

$$X \sim Ber(p) \Rightarrow Z_n \sim Bin(n, p).$$

Remark 1033 *Assume that X is Poisson distributed with rate parameter λ. Then, Z_n is Poisson distributed with rate parameters $n\lambda$. In symbols,*

$$X \sim Poiss(\lambda) \Rightarrow Z_n \sim Poiss(n\lambda).$$

Remark 1034 *Assume that X is Gaussian distributed with mean μ and variance σ^2. Then, Z_n is Gaussian distributed with mean $n\mu$ and variance $n\sigma^2$. In symbols,*

$$X \sim N(\mu, \sigma^2) \Rightarrow Z_n \sim N(n\mu, n\sigma^2).$$

Remark 1035 *Assume that X is exponentially distributed with rate parameter λ. Then, Z_n is gamma distributed with parameters n and λ. In symbols,*

$$X \sim Exp(\lambda) \Rightarrow Z_n \sim \Gamma(n, \lambda).$$

Remark 1036 *Assume that X is exponentially distributed with rate parameter λ. Then, $2\lambda Z_n$ is chi-square distributed with $2n$ degrees of freedom. In symbols,*

$$X \sim Exp(\lambda) \Rightarrow Z_n \sim \frac{1}{2\lambda}\chi^2_{2n}.$$

Remark 1037 *Assume that X is chi-square distributed with one degree of freedom. Then, Z_n in chi-square distributed with n degrees of freedom. In symbols,*

$$X \sim \chi^2_1 \Rightarrow Z_n \sim \chi^2_n.$$

We recall the utmost important central limit theorem (see Theorem 1019), which we restate here in a more statistical language.

Theorem 1038 (central limit theorem) *Assume that X has finite moment of order 2. Write $\mathbf{E}[X] \equiv \mu_X$. and $\mathbf{D}^2[X] \equiv \sigma_X^2$. Then, is the limit as $n \to \infty$, the standardized version of Z_n is standard Gaussian distributed. In symbols,*

$$X \in \mathcal{L}^2(\Omega;\mathbb{R}) \Rightarrow \tilde{Z}_n \xrightarrow{w} N(0,1)$$

where

$$\tilde{Z}_n \overset{def}{=} \frac{Z_n - n\mu_X}{\sqrt{n}\sigma_X}, \quad \forall n \in \mathbb{N}.$$

More explicitely,

$$X \in \mathcal{L}^2(\Omega;\mathbb{R}) \Rightarrow \lim_{n\to\infty} F_{\tilde{Z}_n}(x) = \lim_{n\to\infty} \mathbf{P}\left(\frac{Z_n - n\mu_X}{\sqrt{n}\sigma_X} \leq x\right) = \Phi(x),$$

for every $x \in \mathbb{R}$, where $\Phi : \mathbb{R} \to \mathbb{R}$ is the standard normal distribution function.

Definition 1039 (sample mean) *Consider the Borel function $g : \mathbb{R}^n \to \mathbb{R}$, given by*

$$g(x_1,\ldots,x_n) \overset{def}{=} \frac{1}{n}\sum_{k=1}^{n} x_k, \quad \forall (x_1,\ldots,x_n) \in \mathbb{R}^n. \tag{13.7}$$

We call the statistic $\overline{X}_n \overset{def}{=} g \circ (X_1,\ldots,X_n)$, more commonly written as

$$\bar{X}_n \equiv \frac{1}{n}\sum_{k=1}^{n} X_k, \tag{13.8}$$

the sample mean *of size n drawn from X.*

Proposition 1040 (mean of the sample mean) *Assume that X has finite moment of order 1. Write $\mathbf{E}[X] \equiv \mu_X$. Then, $\bar{X}_n$ has finite moment of order 1. Moreover, we have*

$$\mu_{\bar{X}_n} \equiv \mathbf{E}\left[\bar{X}_n\right] = \mu_X. \tag{13.9}$$

Proof. Observing that

$$\bar{X}_n = \frac{1}{n} Z_n,$$

for every $n \in \mathbb{N}$, the finiteness of the moment of order 1 of $\bar{X}_n$ immediately follows from the finiteness of the moment of order 1 of Z_n (see Proposition 1030). Moreover, thanks to the linearity property of the expectation operator, , on account of Equation (13.5), we can write

$$\mathbf{E}\left[\bar{X}_n\right] = \mathbf{E}\left[\frac{1}{n} Z_n\right] = \frac{1}{n}\mathbf{E}\left[Z_n\right] = \frac{n\mu_X}{n} = \mu_X,$$

as desired. □

Proposition 1041 (variance of the sample mean) *Assume that X has finite moment of order 2. Write $\mathbf{D}^2[X] \equiv \sigma_X^2$. Then, $\bar{X}_n$ has finite moment of order 2. Moreover, we have*

$$\mathbf{D}^2\left[\bar{X}_n\right] \equiv \sigma_{\bar{X}_n}^2 = \frac{\sigma_X^2}{n}. \tag{13.10}$$

Proof. Mutatis mutandis in the first part of the proof of Proposition (1040), we obtain that $\bar{X}_n$ has finite moment of order 2 (see Proposition 1031). Moeover, recalling that the variance operator is square-homogeneous, on account of Equation (13.6), we obtain

$$\mathbf{D}^2\left[\bar{X}_n\right] = \mathbf{D}^2\left[\frac{1}{n} Z_n\right] = \frac{1}{n^2}\mathbf{D}^2\left[Z_n\right] = \frac{n\sigma_X^2}{n^2} = \frac{\sigma_X^2}{n},$$

as desired. □

Corollary 1042 *Assume that X has finite moment of order 1. Write μ. Then, the sample mean $\bar{X}_n$ of size n drawn from X, converges in probability to μ_X as n goes to ∞. In symbols,*

$$\bar{X}_n \xrightarrow{\boldsymbol{P}} \mu_X, \quad \text{as } n \to \infty.$$

Proof. By virtue of the Khintchine Theorem... . □

Corollary 1043 *Assume that X has finite moment of order 2. Then, the sample mean $\bar{X}_n$ of size n drawn from X, converges in mean square to μ_X as n goes to ∞. In symbols,*

$$\bar{X}_n \xrightarrow{\boldsymbol{L}^2} \mu_X, \quad \text{as } n \to \infty.$$

A fortiori, $\bar{X}_n$ converges in probability to μ_X as n goes to ∞.

Proof. In light of Proposition **??**, we have

$$\mathbf{E}\left[\left(\bar{X}_n - \mu_X\right)^2\right] = \mathbf{E}\left[\left(\bar{X}_n - \mathbf{E}\left[\bar{X}_n\right]\right)^2\right] = \mathbf{D}^2\left[\bar{X}_n\right] = \frac{\sigma_X^2}{n}.$$

The desired result clearly follows. □

Proposition 1044 *Assume that X has finite moment of order 3 [resp. 4]. Write $\mu_X^{(3)}$ [resp. $\mu_X^{(4)}$] for the central moment of X of order 3 [resp. 4]. That is $\mu_X^{(3)} \equiv \mathbf{E}\left[(X-\mu_X)^3\right]$ [resp. $\mu_X^{(4)} \equiv \mathbf{E}\left[(X-\mu_X)^4\right]$], where $\mu_X \equiv \mathbf{E}[X]$. Then, we have*

$$\mu_{\bar{X}_n}^{(3)} = \frac{\mu_X^{(3)}}{n^2} \qquad \text{and} \qquad \mu_{\bar{X}_n}^{(4)} = \frac{3\sigma_X^4}{n^2} + \frac{\mu^{(4)} - 3\sigma_X^4}{n^3}$$

where $\sigma_X \equiv \mathbf{D}[X]$.

$$\begin{aligned}\mathbf{M}^{(3)}\left(\bar{X}_n\right) &= \mathbf{E}\left[\left(\bar{X}_n - \mathbf{E}\left[\bar{X}_n\right]\right)^3\right] = \mathbf{E}\left[\left(\frac{1}{n}\sum_{k=1}^n X_k - \mu_X\right)^3\right] \\ &= \frac{1}{n^3}\mathbf{E}\left[\left(\sum_{k=1}^n X_k - n\mu_X\right)^3\right] = \frac{1}{n^3}\mathbf{E}\left[\left(\sum_{k=1}^n (X_k - \mu_X)\right)^3\right].\end{aligned}$$

where

$$\begin{aligned}\left(\sum_{k=1}^n (X_k-\mu_X)\right)^3 &= \sum_{k=1}^n (X_k-\mu_X)^3 + \sum_{\substack{j=1,k=1\\k\neq j}}^n (X_j-\mu_X)(X_k-\mu_X)^2 \\ &\quad + \sum_{\substack{j=1,k=1,\ell=1\\k\neq j,\ell\neq j,\ell\neq k}}^n (X_j-\mu_X)(X_k-\mu_X)(X_\ell-\mu_X).\end{aligned}$$

It follows

$$\begin{aligned}\mathbf{E}\left[\left(\sum_{k=1}^n (X_k-\mu_X)\right)^3\right] &= \sum_{k=1}^n \mathbf{E}\left[(X_k-\mu_X)^3\right] \\ &\quad + \sum_{\substack{j=1,k=1\\k\neq j}}^n \mathbf{E}\left[(X_j-\mu_X)(X_k-\mu_X)^2\right] \\ &\quad + \sum_{\substack{j=1,k=1,\ell=1\\k\neq j,\ell\neq j,\ell\neq k}}^n \mathbf{E}\left[(X_j-\mu_X)(X_k-\mu_X)(X_\ell-\mu_X)\right].\end{aligned}$$

On the other hand, the random veriables of the sample $X_1, \ldots, X_n$ are independent and have the same distribution as X. We Then, obtain

$$\sum_{k=1}^n \mathbf{E}\left[(X_k-\mu_X)^3\right] = \sum_{k=1}^n \mathbf{E}\left[(X-\mu_X)^3\right] = \sum_{k=1}^n \mu_X^{(3)} = n\mu_X^{(3)}$$

$$\begin{aligned}\sum_{\substack{j=1,k=1\\k\neq j}}^n \mathbf{E}\left[(X_j-\mu_X)(X_k-\mu_X)^2\right] &= \sum_{\substack{j=1,k=1\\k\neq j}}^n \mathbf{E}[X_j-\mu_X]\,\mathbf{E}\left[(X_k-\mu_X)^2\right] \\ &= \sum_{\substack{j=1,k=1\\k\neq j}}^n \mathbf{E}[X-\mu_X]\,\mathbf{E}\left[(X-\mu_X)^2\right] \\ &= 0\end{aligned}$$

$$\begin{aligned}\sum_{\substack{j=1,k=1,\ell=1\\k\neq j,\ell\neq j,\ell\neq k}}^{n}\mathbf{E}\left[\left(X_j-\mu_X\right)\left(X_k-\mu_X\right)\left(X_\ell-\mu_X\right)\right] &= \sum_{\substack{j=1,k=1,\ell=1\\k\neq j,\ell\neq j,\ell\neq k}}^{n}\mathbf{E}\left[\left(X_j-\mu_X\right)\right]\mathbf{E}\left[\left(X_\ell-\mu_X\right)\right]\mathbf{E}\left[\left(X_\ell-\mu_X\right)\right]\\ &= \sum_{\substack{j=1,k=1,\ell=1\\k\neq j,\ell\neq j,\ell\neq k}}^{n}\mathbf{E}\left[\left(X-\mu_X\right)\right]\mathbf{E}\left[\left(X-\mu_X\right)\right]\mathbf{E}\left[\left(X-\mu_X\right)\right]\\ &= 0\end{aligned}$$

Proposition 1045 *Assume that X is Gaussian distributed with mean μ_X and variance σ_X^2. Then, $\bar{X}_n$ is normally distributed with mean μ_X and variance σ_X^2/n. In symbols,*

$$X\sim N\left(\mu_X,\sigma_X^2\right)\Rightarrow \bar{X}_n\sim N\left(\mu_X,\sigma_X^2/n\right).$$

In addition, we have

$$\mu_{\bar{X}_n}^{(2n-1)}=0$$

for every $n\in\mathbb{N}$. Furthermore,

$$\mu_{\bar{X}_n}^{(4)}=\frac{3\sigma_X^4}{n^2}\qquad \textit{and}\qquad \mu_{\bar{X}_n}^{(6)}=\frac{15\sigma_X^6}{n^3}$$

Proof. See [3, p. 346]. □

Remark 1046 *Assume that X has finite mean μ and variance σ^2. Then, in the limit as $n\to\infty$, the standardized version of $\bar{X}_n$, that is the random variable $\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}$, has the standard normal distribution. In symbols,*

$$X\in\mathcal{L}^2\left(\Omega;\mathbb{R}\right)\Rightarrow\lim_{n\to\infty}\mathbf{P}\left(\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}\leq x\right)=\Phi\left(x\right),\quad\forall x\in\mathbb{R},$$

where $\Phi:\mathbb{R}\to\mathbb{R}$ is the standard normal distribution function.

We recall also two Laws of Large Numbers (see Corollary 1007 and Theorem 1016), which are the most frequently applied in Statistics.

Theorem 1047 *Assume that X has finite mean μ and variance σ^2. Then, in the limit as $n\to\infty$, the sample mean $\bar{X}_n$ converges to μ in probability and in mean square. In symbols,*

$$X\in\mathcal{L}^2\left(\Omega;\mathbb{R}\right)\Rightarrow\begin{cases}\lim_{n\to\infty}\mathbf{P}\left(\left|\bar{X}_n-\mu\right|\geq\varepsilon\right)=0, & \forall\varepsilon>0,\\ \lim_{n\to\infty}\mathbf{E}\left[\left(\bar{X}_n-\mu\right)^2\right]=0.\end{cases}$$

Definition 1048 *Consider the Borel function $g:\mathbb{R}^n\to\mathbb{R}$, given by*

$$g\left(x_1,\ldots,x_n\right)\overset{def}{=}\max\left(x_1,\ldots,x_n\right),\quad\forall\left(x_1,\ldots,x_n\right)\in\mathbb{R}^n,$$

where $\max\left(x_1,\ldots,x_n\right)$ is the maximum of the n-tuple of real numbers $\left(x_1,\ldots,x_n\right)$. We call the statistic $\check{X}_n\overset{def}{=}g\circ\left(X_1,\ldots,X_n\right)$, written also as $\max\left(X_1,\ldots,X_n\right)$ or $\bigvee_{k=1}^n X_k$, the sample maximum *of size n drawn from X (see Example 451).*

Proposition 1049 *Assume that X is exponentially distributed with rate parameter λ. Then, the ditribution function $F_{\check{X}_n} : \mathbb{R} \to \mathbb{R}$ of $\check{X}_n$ is given by*

$$F_{\check{X}_n}(x) = \left(1 - e^{-\lambda x}\right)^n 1_{\mathbb{R}_+}(x) \tag{13.11}$$

for every $x \in \mathbb{R}$.

Proof. In terms of events, we clearly have

$$\left\{\check{X}_n \leq x\right\} = \{X_1 \leq x, \ldots, X_n \leq x\},$$

for every $x \in \mathbb{R}$. Therefore, since $X_1, \ldots, X_n$ are independent and have the same distribution of X

$$\begin{aligned} F_{\check{X}_n}(x) &= \mathbf{P}\left(\check{X}_n \leq x\right) = \mathbf{P}(X_1 \leq x, \ldots, X_n \leq x) = \textstyle\prod_{k=1}^n \mathbf{P}(X_k \leq x) \\ &= \textstyle\prod_{k=1}^n \mathbf{P}(X \leq x) = \mathbf{P}(X \leq x)^n = \left(1 - e^{-\lambda x}\right)^n 1_{\mathbb{R}_+}(x), \end{aligned}$$

for every $x \in \mathbb{R}$. □

Proposition 1050 *Assume that X is exponentially distributed with rate parameter λ. We have*

$$\mathbf{E}\left[\check{X}_n\right] = \frac{1}{\lambda} \textstyle\sum_{k=1}^n \frac{1}{k}. \tag{13.12}$$

Proof. The derivative of the distribution function $F_{\check{X}_n} : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}_+$ is given by

$$F'_{\check{X}_n}(x) = n\lambda \left(1 - e^{-\lambda x}\right)^{n-1} e^{-\lambda x} 1_{\mathbb{R}_+}(x),$$

for every $x \in \mathbb{R}$ and every $\lambda \in \mathbb{R}_+$. Moreover, it is clearly seen that $F'_{\check{X}_n}$ is bounded for every fixed $\lambda \in \mathbb{R}_+$. Hence, $\check{X}_n$ is absolutely continuous with density $f_{\check{X}_n} : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}_+$ given by

$$f_{\check{X}_n}(x) = F'_{\check{X}_n}(x),$$

for every $x \in \mathbb{R}$ and every $\lambda \in \mathbb{R}_+$. It follows

$$\begin{aligned} \mathbf{E}\left[\check{X}_n\right] &= \int_{\mathbb{R}} x f_{\check{X}_n}(x)\, d\mu_L(x) = \int_{\mathbb{R}} xn\lambda \left(1 - e^{-\lambda x}\right)^{n-1} e^{-\lambda x} 1_{\mathbb{R}_+}(x)\, d\mu_L(x) \\ &= \int_{\mathbb{R}_+} n\lambda x \left(1 - e^{-\lambda x}\right)^{n-1} e^{-\lambda x} d\mu_L(x) = \int_0^{+\infty} n\lambda x \left(1 - e^{-\lambda x}\right)^{n-1} e^{-\lambda x} dx. \end{aligned} \tag{13.13}$$

On the other hand,

$$\int_0^{+\infty} n\lambda x \left(1 - e^{-\lambda x}\right)^{n-1} e^{-\lambda x} dx = \int_0^{+\infty} 1 - \left(1 - e^{-\lambda x}\right)^n dx. \tag{13.14}$$

In fact, since

$$\lim_{x \to +\infty} \left(1 - \left(1 - e^{-\lambda x}\right)^n\right) x = \lim_{x \to +\infty} \left(1 - \textstyle\sum_{k=0}^n (-1)^k \binom{n}{k} e^{-k\lambda x}\right) x = \textstyle\sum_{k=1}^n (-1)^k \binom{n}{k} x e^{-k\lambda x} = \textstyle\sum_{k=1}^n (-1)^k$$

integrating by parts, we obtain

$$\begin{aligned}\int_0^{+\infty} 1-\left(1-e^{-\lambda x}\right)^n dx &= \left(1-\left(1-e^{-\lambda x}\right)^n\right) x\Big|_0^{+\infty} - \int_0^{+\infty} x d\left(1-\left(1-e^{-\lambda x}\right)^n\right) \\ &= \int_0^{+\infty} n\lambda x \left(1-e^{-\lambda x}\right)^{n-1} e^{-\lambda x} dx.\end{aligned}$$

Now,

$$\begin{aligned}\int_0^{+\infty} 1-\left(1-e^{-\lambda x}\right)^n dx &= \int_0^{+\infty} \left(1-\textstyle\sum_{k=0}^n (-1)^k \binom{n}{k} e^{-k\lambda x}\right) dx \\ &= \int_0^{+\infty} \textstyle\sum_{k=1}^n (-1)^{k+1} \binom{n}{k} e^{-k\lambda x} dx \\ &= \textstyle\sum_{k=1}^n (-1)^{k+1} \binom{n}{k} \int_0^{+\infty} e^{-k\lambda x} dx \\ &= \frac{1}{\lambda} \textstyle\sum_{k=1}^n (-1)^k \binom{n}{k} \frac{1}{k} \int_0^{+\infty} e^{-k\lambda x} d(-k\lambda x) \\ &= \frac{1}{\lambda} \textstyle\sum_{k=1}^n (-1)^k \binom{n}{k} \frac{1}{k} \int_0^{-\infty} e^{u} du \\ &= \frac{1}{\lambda} \textstyle\sum_{k=1}^n (-1)^{k+1} \binom{n}{k} \frac{1}{k} \int_{-\infty}^0 e^{u} du \\ &= \frac{1}{\lambda} \textstyle\sum_{k=1}^n (-1)^{k+1} \binom{n}{k} \frac{1}{k}. \end{aligned} \tag{13.15}$$

Furthermore,

$$\begin{aligned}\textstyle\sum_{k=1}^n (-1)^{k+1} \binom{n}{k} \frac{1}{k} &= \textstyle\sum_{k=1}^n (-1)^{k+1} \binom{n}{k} \int_0^1 x^{k-1} dx \\ &= \int_0^1 -\textstyle\sum_{k=1}^n (-1)^k \binom{n}{k} x^{k-1} dx \\ &= \int_0^1 \frac{1-(1-x)^n}{x} dx \\ &= \int_0^1 \frac{1-y^n}{1-y} dy \\ &= \int_0^1 \sum_{k=0}^{n-1} y^k dy \\ &= \sum_{k=0}^{n-1} \int_0^1 y^k dy \\ &= \sum_{k=0}^{n-1} \frac{1}{k+1} y^{k+1}\Big|_0^1 \\ &= \sum_{k=1}^{n} \frac{1}{k}. \end{aligned} \tag{13.16}$$

In the end, combining (13.13)-(13.16), the desired (13.12) follows.**Proof.**

Definition 1051 *Consider the Borel function $g : \mathbb{R}^n \to \mathbb{R}$, given by*

$$g(x_1, \ldots, x_n) \stackrel{def}{=} \min(x_1, \ldots, x_n), \quad \forall (x_1, \ldots, x_n) \in \mathbb{R}^n,$$

where $\min(x_1, \ldots, x_n)$ is the minimum of the n-tuple of real numbers $(x_1, \ldots, x_n)$. We call the statistic $\hat{X}_n \stackrel{def}{=} g \circ (X_1, \ldots, X_n)$, written also as $\min(X_1, \ldots, X_n)$ or $\bigwedge_{k=1}^n X_k$, the sample minimum *of size n drawn from X (see Example 452).*

Proposition 1052 *Assume that X is exponentially distributed with rate parameter λ. Then, $\hat{X}_n$ is exponentially distributed with rate parameter $n\lambda$. In symbols*

$$X \sim Exp(\lambda) \Rightarrow \hat{X}_n \sim Exp(n\lambda).$$

Proof. In terms of events, we clearly have

$$\left\{\hat{X}_n \leq x\right\}^c = \left\{\hat{X}_n > x\right\} = \{X_1 > x, \ldots, X_n > x\}$$

for every $x \in \mathbb{R}$. On the other hand, since the random variables $X_1, \ldots, X_n$ are independent and have the same distribution of X,

$$\begin{aligned}\mathbf{P}(X_1 > x, \ldots, X_n > x) &= \textstyle\prod_{k=1}^n \mathbf{P}(X_k > x) = \prod_{k=1}^n \mathbf{P}(X > x) = \mathbf{P}(X > x)^n \\ &= (1 - \mathbf{P}(X \leq x))^n = \left(1 - \left(1 - e^{-\lambda x}\right) 1_{\mathbb{R}_+}(x)\right)^n = 1 - \left(1 - e^{-n\lambda x}\right) 1_{\mathbb{R}_+}(x).\end{aligned}$$

Therefore, the distribution function $F_{\hat{X}_n} : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}_+$ of $\hat{X}_n$ is given by

$$F_{\hat{X}_n}(x) = \mathbf{P}\left(\hat{X}_n \leq x\right) = 1 - \mathbf{P}\left(\left\{\hat{X}_n \leq x\right\}^c\right) = 1 - \mathbf{P}(X_1 > x, \ldots, X_n > x) = \left(1 - e^{-n\lambda x}\right) 1_{\mathbb{R}_+}(x),$$

for every $x \in \mathbb{R}$, and every $\lambda \in \mathbb{R}_+$. □

Definition 1053 *Consider the Borel function $g : \mathbb{R}^n \to \mathbb{R}$, given by*

$$g(x_1, \ldots, x_n) \stackrel{def}{=} \operatorname{med}(x_1, \ldots, x_n), \quad \forall (x_1, \ldots, x_n) \in \mathbb{R}^n,$$

*where $\operatorname{med}(x_1, \ldots, x_n)$ is the median of the n-tuple of real numbers $(x_1, \ldots, x_n)$ (see Definition **??**). We call the statistic $\tilde{X}_n \stackrel{def}{=} g \circ (X_1, \ldots, X_n)$, more commonly written as $\operatorname{med}(X_1, \ldots, X_n)$, the* sample median *of size n drawn from X.*

Definition 1054 *Given any integer m such that $0 \leq m < 50$, consider the Borel function $g : \mathbb{R}^n \to \mathbb{R}$, given by*

$$g(x_1, \ldots, x_n) \stackrel{def}{=} \operatorname{mean}_{\mathbf{tr}(m)}(x_1, \ldots, x_n), \quad \forall (x_1, \ldots, x_n) \in \mathbb{R}^n,$$

*where $\operatorname{mean}_{\mathbf{tr}(m)}(x_1, \ldots, x_n)$ is the $m\%$ trimmed mean of the n-tuple of real numbers $(x_1, \ldots, x_n)$ (see Definition **??**). We call the statistic $\bar{X}_{n,\mathbf{tr}(m)} \stackrel{def}{=} g \circ (X_1, \ldots, X_n)$, more commonly written as $\operatorname{mean}_{\mathbf{tr}(m)}(X_1, \ldots, X_n)$, the* sample $m\%$ trimmed mean *of size n drawn from X.*

Definition 1055 *Consider the Borel function $g : \mathbb{R}^n \to \mathbb{R}$, given by*

$$g(x_1, \ldots, x_n) \stackrel{def}{=} \text{midrange}(x_1, \ldots, x_n), \quad \forall (x_1, \ldots, x_n) \in \mathbb{R}^n,$$

where midrange $(x_1, \ldots, x_n)$ *is the midrange of the n-tuple of real numbers $(x_1, \ldots, x_n)$*[1]*. We call the statistic $\bar{X}_{n,\mathbf{e}} \stackrel{def}{=} g \circ (X_1, \ldots, X_n)$, more commonly written as* midrange $(X_1, \ldots, X_n)$*, the* sample midrange *of size n drawn from X.*

Definition 1056 *Assume that X has finite mean μ and variance σ^2 and consider the Borel function $g : \mathbb{R}^n \to \mathbb{R}$ given by*

$$g(x_1, \ldots, x_n) \stackrel{def}{=} \sum_{k=1}^{n} x_k^2, \quad \forall (x_1, \ldots, x_n) \in \mathbb{R}^n. \tag{13.17}$$

We call the statistic $Q_n \stackrel{def}{=} g \circ (X_1, \ldots, X_n)$, more commonly written as $\sum_{k=1}^{n} X_k^2$, the sample square sum *of size n drawn from X.*

Lemma 1057 *Assume that X is standard Gaussian distributed. Then, X^2 has a chi-square distribution with one degree of freedom. In symbols,*

$$X \sim N(0,1) \Rightarrow X^2 \sim \chi_1^2.$$

Proof. . □

Remark 1058 *Assume that X is standard Gaussian distributed. Then, Q_n has a chi-square distribution with n degree of freedom. In symbols,*

$$X \sim N(0,1) \Rightarrow Q_n \sim \chi_n^2.$$

Proof. . □

Definition 1059 *Assume that X has finite moment of order 1 and consider the Borel function $g : \mathbb{R}^n \to \mathbb{R}$, given by*

$$g(x_1, \ldots, x_n) \stackrel{def}{=} \frac{1}{n} \sum_{k=1}^{n} (x_k - \mu_X)^2, \quad \forall (x_1, \ldots, x_n) \in \mathbb{R}^n. \tag{13.18}$$

where $\mu_X \equiv \mathbf{E}[X]$. We call the statistic

$$S^2_{X,\mu_X,n} \stackrel{def}{=} g \circ (X_1, \ldots, X_n) \equiv \frac{1}{n} \sum_{k=1}^{n} (X_k - \mu_X)^2 \tag{13.19}$$

the sample variance about the mean *of size n drawn from X.*

[1] Recall that for any n-tuple of real numbers $(x_1, \ldots, x_n)$ the midrange is given by

$$\text{midrange}(x_1, \ldots, x_n) \stackrel{\text{def}}{=} \frac{1}{2} (\min(x_1, \ldots, x_n) + \max(x_1, \ldots, x_n)).$$

Proposition 1060 *Assume that X has finite moment of order 2. We have*

$$\mathbf{E}\left[S^2_{X,\mu_X,n}\right] = \sigma^2_X. \tag{13.20}$$

where $\sigma^2_X \equiv \mathbf{D}^2[X]$.

Proof. In fact,

$$\mathbf{E}\left[S^2_{X,\mu_X,n}\right] = \frac{1}{n}\sum_{k=1}^{n}\mathbf{E}\left[(X_k-\mu_X)^2\right] = \frac{1}{n}\sum_{k=1}^{n}\mathbf{E}\left[(X_k-\mu_X)^2\right] = \frac{1}{n}\sum_{k=1}^{n}\mathbf{E}\left[(X-\mu_X)^2\right] = \frac{1}{n}\sum_{k=1}^{n}\sigma^2_X = \sigma^2_X.$$

as claimed. □

Proposition 1061 *Under the further assumption that X has finite central moment of the fourth order, μ_4, we have*

$$\mathbf{D}^2\left[S^2_n(\mu)\right] = \frac{1}{n}\left(\mu^{(4)}_X - \sigma^4_X\right). \tag{13.21}$$

Proof. A straightforward computation yields

$$\begin{aligned}\mathbf{D}^2\left[S^2_n(\mu)\right] &= \mathbf{D}^2\left[\frac{1}{n}\sum_{k=1}^{n}(X_k-\mu_X)^2\right] = \frac{1}{n^2}\sum_{k=1}^{n}\mathbf{D}^2\left[(X_k-\mu_X)^2\right] = \frac{1}{n^2}\sum_{k=1}^{n}\mathbf{D}^2\left[(X-\mu_X)^2\right]\\ &= \frac{1}{n^2}\sum_{k=1}^{n}\left(\mathbf{E}\left[(X-\mu_X)^4\right] - \mathbf{E}\left[(X-\mu_X)^2\right]^2\right) = \frac{1}{n^2}\sum_{k=1}^{n}\left(\mu^{(4)}_X - \sigma^4_X\right)\\ &= \frac{1}{n}\left(\mu^{(4)}_X - \sigma^4_X\right),\end{aligned}$$

as claimed. □

Definition 1062 (unbiased sample variance) *Consider the Borel function $g : \mathbb{R}^n \to \mathbb{R}$, given by*

$$g(x_1,\ldots,x_n) \stackrel{def}{=} \frac{1}{n-1}\sum_{k=1}^{n}\left(x_k - \frac{1}{n}\sum_{\ell=1}^{n}x_\ell\right)^2, \quad \forall\,(x_1,\ldots,x_n)\in\mathbb{R}^n. \tag{13.22}$$

We call the statistic

$$S^2_{X,n} \stackrel{def}{=} g\circ(X_1,\ldots,X_n) \equiv \frac{1}{n-1}\sum_{k=1}^{n}\left(X_k-\bar{X}_n\right)^2 \tag{13.23}$$

the unbiased sample variance *of size n drawn from X.*

Lemma 1063 (unbiased sample variance) *We can write*

$$S^2_{X,n} = \frac{1}{n-1}\left(\sum_{k=1}^{n}(X_k-\mu_X)^2 - \frac{1}{n}\left(\sum_{k=1}^{n}(X_k-\mu_X)\right)^2\right), \tag{13.24}$$

for every $\mu\in\mathbb{R}$.

Proof. In fact,

$$\begin{aligned}
S_{X,n}^2 &= \frac{1}{n-1}\sum_{k=1}^{n}\left(X_k - \frac{1}{n}\sum_{\ell=1}^{n} X_\ell\right)^2 \\
&= \frac{1}{n-1}\sum_{k=1}^{n}\left(X_k - \mu_X + \mu_X - \frac{1}{n}\sum_{\ell=1}^{n} X_\ell\right)^2 \\
&= \frac{1}{n-1}\sum_{k=1}^{n}\left((X_k - \mu_X) - \frac{1}{n}\sum_{\ell=1}^{n} (X_\ell - \mu_X)\right)^2 \\
&= \frac{1}{n-1}\sum_{k=1}^{n}\left((X_k - \mu_X)^2 - \frac{2}{n}(X_k - \mu_X)\sum_{\ell=1}^{n}(X_\ell - \mu_X) + \frac{1}{n^2}\left(\sum_{\ell=1}^{n}(X_\ell - \mu_X)\right)^2\right) \\
&= \frac{1}{n-1}\left(\sum_{k=1}^{n}(X_k - \mu_X)^2 - \frac{2}{n}\sum_{k=1}^{n}\left((X_k - \mu_X)\sum_{\ell=1}^{n}(X_\ell - \mu_X)\right) + \frac{1}{n^2}\sum_{k=1}^{n}\left(\sum_{\ell=1}^{n}(X_\ell - \mu_X)\right)^2\right) \\
&= \frac{1}{n-1}\left(\sum_{k=1}^{n}(X_k - \mu_X)^2 - \frac{2}{n}\left(\sum_{k=1}^{n}(X_k - \mu_X)\right)\left(\sum_{\ell=1}^{n}(X_\ell - \mu_X)\right) + \frac{1}{n^2}n\left(\sum_{\ell=1}^{n}(X_\ell - \mu_X)\right)^2\right) \\
&= \frac{1}{n-1}\left(\sum_{k=1}^{n}(X_k - \mu_X)^2 - \frac{2}{n}\left(\sum_{k=1}^{n}(X_k - \mu_X)\right)^2 + \frac{1}{n}\left(\sum_{k=1}^{n}(X_k - \mu_X)\right)^2\right) \\
&= \frac{1}{n-1}\left(\sum_{k=1}^{n}(X_k - \mu_X)^2 - \frac{1}{n}\left(\sum_{k=1}^{n}(X_k - \mu_X)\right)^2\right),
\end{aligned}$$

as desired. □

Proposition 1064 *Assume that X has finite mean μ_X and variance σ_X^2. We have*

$$\mathbf{E}\left[S_{X,n}^2\right] = \sigma_X^2. \tag{13.25}$$

Proof. Thanks to (13.24), we can write

$$\begin{aligned}
\mathbf{E}\left[S_{X,n}^2\right] &= \mathbf{E}\left[\frac{1}{n-1}\left(\sum_{k=1}^{n}(X_k - \mu_X)^2 - \frac{1}{n}\left(\sum_{k=1}^{n}(X_k - \mu_X)\right)^2\right)\right] \\
&= \frac{1}{n-1}\left(\sum_{k=1}^{n}\mathbf{E}\left[(X_k - \mu_X)^2\right] - \frac{1}{n}\mathbf{E}\left[\left(\sum_{k=1}^{n}(X_k - \mu_X)\right)^2\right]\right).
\end{aligned} \tag{13.26}$$

Now,

$$\sum_{k=1}^{n}\mathbf{E}\left[(X_k - \mu_X)^2\right] = n\sigma_X^2. \tag{13.27}$$

On the other hand, on account of the independence of the random variables $X_1, \ldots, X_n$,

$$\begin{aligned}
\mathbf{E}\left[\left(\sum_{k=1}^{n}(X_k-\mu_X)\right)^2\right] &= \mathbf{E}\left[\sum_{k=1}^{n}(X_k-\mu_X)^2+\sum_{\substack{k,\ell=1\\ \ell\neq k}}^{n}(X_k-\mu_X)(X_\ell-\mu_X)\right] \\
&= \sum_{k=1}^{n}\mathbf{E}\left[(X_k-\mu_X)^2\right]+\sum_{\substack{k,\ell=1\\ \ell\neq k}}^{n}\mathbf{E}\left[X_k-\mu_X\right]\mathbf{E}\left[X_\ell-\mu_X\right] \\
&\sum_{k=1}^{n}\mathbf{E}\left[(X_k-\mu_X)^2\right]+\sum_{\substack{k,\ell=1\\ \ell\neq k}}^{n}\left(\mathbf{E}\left[X_k\right]-\mu_X\right)\left(\mathbf{E}\left[X_\ell\right]-\mu_X\right) \\
&= n\sigma_X^2. \qquad (13.28)
\end{aligned}$$

Replacing (13.27) and (13.28) into (13.26), the desired result immediately follows. □

Proposition 1065 *Under the further assumption that X has finite moment of order four, we have*

$$\mathbf{D}^2\left[S_{X,n}^2\right]=\frac{1}{n}\left(\mu_X^{(4)}-\frac{n-3}{n-1}\sigma_X^4\right)=\frac{\sigma_X^4}{n}\left(\kappa-\frac{n-3}{n-1}\right), \qquad (13.29)$$

where $\mu_X^{(4)}$ [resp. $\kappa\equiv\hat{\mu}_X^{(4)}\equiv\mu_X^{(4)}/\sigma_X^4$] is the central moment of order four [resp. kurtosis or standardized central moment of order four] of X.

Proof. Assuming that X has finite central moment of the fourth order, we can write

$$\mathbf{D}^2\left[S_{X,n}^2\right]=\mathbf{E}\left[\left(S_{X,n}^2\right)^2\right]-\mathbf{E}\left[S_{X,n}^2\right]^2. \qquad (13.30)$$

By virtue of **??** of Proposition 1064, we know that

$$\mathbf{E}\left[S_{X,n}^2\right]^2=\sigma_X^4.$$

Therfore, we are left with computing $\mathbf{E}\left[\left(S_{X,n}^2\right)^2\right]$. On the other hand, thanks to (**??**) of Lemma **??**, we can write

$$\begin{aligned}
\left(S_{X,n}^2\right)^2 &= \left(\frac{1}{n-1}\left(\sum_{k=1}^{n}(X_k-\mu_X)^2-\frac{1}{n}\left(\sum_{k=1}^{n}(X_k-\mu_X)\right)^2\right)\right)^2 \\
&= \frac{1}{n^2(n-1)^2}\left(n^2\left(\sum_{k=1}^{n}(X_k-\mu_X)^2\right)^2-2n\left(\sum_{k=1}^{n}(X_k-\mu_X)^2\right)\left(\sum_{k=1}^{n}(X_k-\mu_X)\right)^2+\left(\sum_{k=1}^{n}(X_k-\mu\right.\right. \qquad (13.31)
\end{aligned}$$

Now, we have

$$\left(\sum_{k=1}^{n}(X_k-\mu_X)^2\right)^2=\sum_{k=1}^{n}(X_k-\mu_X)^4+\sum_{\substack{k,\ell=1\\ \ell\neq k}}^{n}(X_k-\mu_X)^2(X_\ell-\mu_X)^2 \qquad (13.32)$$

and

$$\begin{aligned}
&\left(\sum_{k=1}^{n}(X_k-\mu_X)^2\right)\left(\sum_{k=1}^{n}(X_k-\mu_X)\right)^2\\
&=\left(\sum_{k=1}^{n}(X_k-\mu_X)^2\right)\left(\sum_{k=1}^{n}(X_k-\mu_X)^2+\sum_{\substack{k,\ell=1\\ \ell\neq k}}^{n}(X_k-\mu_X)(X_\ell-\mu_X)\right)\\
&=\left(\sum_{k=1}^{n}(X_k-\mu_X)^2\right)^2+\left(\sum_{k=1}^{n}(X_k-\mu_X)^2\right)\sum_{\substack{k,\ell=1\\ \ell\neq k}}^{n}(X_k-\mu_X)(X_\ell-\mu_X)\\
&=\left(\sum_{k=1}^{n}(X_k-\mu_X)^2\right)^2+\sum_{\substack{j,k,\ell=1\\ \ell\neq k,k=j}}^{n}(X_j-\mu_X)^2(X_k-\mu_X)(X_\ell-\mu_X)\\
&+\sum_{\substack{k,j,\ell=1\\ \ell\neq k,\ell=j}}^{n}(X_j-\mu_X)^2(X_k-\mu_X)(X_\ell-\mu_X)+\sum_{\substack{k,j,\ell=1\\ \ell\neq k,k\neq j,\ell\neq j}}^{n}(X_j-\mu_X)^2(X_k-\mu_X)(X_\ell-\mu_X)
\end{aligned}\tag{13.33}$$

and

$$\begin{aligned}
&\left(\sum_{k=1}^{n}(X_k-\mu_X)\right)^4\\
&=\sum_{k=1}^{n}(X_k-\mu_X)^4+\binom{4}{3}\sum_{\substack{k,\ell=1\\ \ell\neq k}}^{n}(X_k-\mu_X)^3(X_\ell-\mu_X)+\frac{1}{2}\binom{4}{2}\sum_{\substack{k,\ell=1\\ \ell\neq k}}^{n}(X_k-\mu_X)^2(X_\ell-\mu_X)^2\\
&+\binom{4}{3}\sum_{\substack{j,k,\ell=1\\ \ell\neq k,\ell\neq j,k\neq j}}^{n}(X_j-\mu_X)^2(X_k-\mu_X)(X_\ell-\mu_X)+\sum_{\substack{i,j,k,\ell=1\\ \ell\neq k,\ell\neq j,\ell\neq i,k\neq j,k\neq i,j\neq i,}}^{n}(X_i-\mu_X)(X_j-\mu_X)(X_k-\mu_X)(X
\end{aligned}\tag{13.34}$$

Hence, from (13.32)-(13.34), on account of the independence of the random variables $X_1,\ldots,X_n$, we obtain

$$\mathbf{E}\left[\left(\sum_{k=1}^{n}(X_k-\mu_X)^2\right)^2\right]=\sum_{k=1}^{n}\mathbf{E}\left[(X_k-\mu_X)^4\right]+\sum_{\substack{k,\ell=1\\ \ell\neq k}}^{n}\mathbf{E}\left[(X_k-\mu_X)^2\right]\mathbf{E}\left[(X_\ell-\mu_X)^2\right]=n\mu_X^{(4)}+n(n-1)\sigma_X^4\tag{13.35}$$

and

$$\mathbf{E}\left[\left(\sum_{k=1}^{n}(X_k-\mu_X)^2\right)\left(\sum_{\ell=1}^{n}(X_\ell-\mu_X)\right)^2\right]=\mathbf{E}\left[\left(\sum_{k=1}^{n}(X_k-\mu_X)^2\right)^2\right]=n\mu_X^{(4)}+n(n-1)\sigma_X^4\tag{13.36}$$

and

$$\mathbf{E}\left[\left(\sum_{k=1}^{n}\left(X_\ell-\mu_X\right)\right)^4\right]=\sum_{k=1}^{n}\mathbf{E}\left[\left(X_k-\mu_X\right)^4\right]+3\sum_{\substack{k,\ell=1\\k\neq\ell}}^{n}\mathbf{E}\left[\left(X_k-\mu_X\right)^2\right]\mathbf{E}\left[\left(X_\ell-\mu_X\right)^2\right]=n\mu_X^{(4)}+3n\left(n-1\right)\sigma \tag{13.37}$$

As a consequence, computing the expectation of both sides of (13.31) and replacing (13.34)-(13.37), it follows

$$\begin{aligned}
&\mathbf{E}\left[\left(S_{X,n}^2\right)^2\right]\\
&=\frac{1}{n^2\left(n-1\right)^2}\left(n^2\mathbf{E}\left[\left(\sum_{k=1}^{n}\left(X_k-\mu\right)^2\right)^2\right]-2n\mathbf{E}\left[\left(\sum_{k=1}^{n}\left(X_k-\mu\right)^2\right)\left(\sum_{k=1}^{n}\left(X_\ell-\mu\right)\right)^2\right]+\mathbf{E}\left[\left(\sum_{k=1}^{n}\left(X_k-\right.\right.\right.\right.\\
&=\frac{1}{n^2\left(n-1\right)^2}\left(n^2\left(n\mu_X^{(4)}+3n\left(n-1\right)\sigma_X^4\right)-2n\left(n\mu_X^{(4)}+3n\left(n-1\right)\sigma_X^4\right)+n\mu_X^{(4)}+3n\left(n-1\right)\sigma_X^4\right)\\
&=\frac{1}{n}\mu_X^{(4)}+\frac{n^2-2n+3}{n\left(n-1\right)}\sigma_X^4.
\end{aligned}$$

In the end,

$$\mathbf{D}^2\left[\left(S_n^2\right)^2\right]=\frac{1}{n}\mu_X^{(4)}+\frac{n^2-2n+3}{n\left(n-1\right)}\sigma_X^4-\sigma_X^4=\frac{1}{n}\left(\mu_X^{(4)}-\frac{n-3}{n-1}\sigma_X^4\right),$$

as claimed. □

Definition 1066 *Consider the Borel function $g:\mathbb{R}^n\to\mathbb{R}$, given by*

$$g\left(x_1,\ldots,x_n\right)\overset{def}{=}\frac{1}{n}\sum_{k=1}^{n}\left(x_k-\frac{1}{n}\sum_{\ell=1}^{n}x_\ell\right)^2,\quad\forall\left(x_1,\ldots,x_n\right)\in\mathbb{R}^n. \tag{13.38}$$

We call the statistic $\tilde{S}_n^2:\nleqslant\to\mathbb{R}$ given by

$$\tilde{S}_n^2\overset{def}{=}g\circ\left(X_1,\ldots,X_n\right)\equiv\frac{1}{n}\sum_{k=1}^{n}\left(X_k-\bar{X}_n\right)^2 \tag{13.39}$$

the biased sample variance *of size n drawn from X.*

Proposition 1067 *Assume that X has finite mean μ and variance σ^2. We have*

$$\tilde{S}_{X,n}^2=\frac{n-1}{n}S_{X,n}^2 \tag{13.40}$$

and

$$\mathbf{E}\left[\tilde{S}_{X,n}^2\right]=\frac{n-1}{n}\sigma^2. \tag{13.41}$$

Proof. By definition,

$$\tilde{S}_{X,n}^2=\frac{n-1}{n}S_{X,n}^2.$$

Therefore, on account of Proposition (**??**), the first-order homogeneous property of the expectation operator yields

$$\mathbf{E}\left[\tilde{S}^2_{X,n}\right] = \mathbf{E}\left[\frac{n-1}{n}S^2_{X,n}\right] = \frac{n-1}{n}\mathbf{E}\left[S^2_{X,n}\right] = \frac{n-1}{n}\sigma^2.$$

as claimed. □

Theorem 1068 *Assume that X is normally distributed with mean μ_X and variance σ^2_X. Then, the statistics $\bar{X}_n$ and $S^2_{X,n}$ are independent.*

Proof. Since

$$\sum_{k=1}^{n}\left(X_k - \bar{X}_n\right) = \sum_{k=1}^{n}\left(X_k - \frac{1}{n}Z_n\right) = \sum_{k=1}^{n}X_k - \frac{1}{n}\sum_{k=1}^{n}Z_n = \sum_{k=1}^{n}X_k - Z_n = 0,$$

we can write

$$X_1 - \bar{X}_n = -\sum_{k=2}^{n}\left(X_k - \bar{X}_n\right).$$

As a consequence, according to **??** of Definition 1062 we obtain

$$\begin{aligned} S^2_{X,n} &= \frac{1}{n-1}\left(\left(X_1 - \bar{X}_n\right)^2 + \sum_{k=2}^{n}\left(X_k - \bar{X}_n\right)^2\right) \\ &= \frac{1}{n-1}\left(\left(\sum_{k=2}^{n}\left(X_k - \bar{X}_n\right)\right)^2 + \sum_{k=2}^{n}\left(X_k - \bar{X}_n\right)^2\right). \end{aligned}$$

Hence, $S^2_{X,n}$ can be represented only in terms of the random variables $X_k - \bar{X}_n$, for $k = 2, \ldots, n$. Therefore, the desired result will follow once we show that the random variable $Y_1 \equiv \bar{X}_n$ is independent of $Y_2 \equiv X_2 - \bar{X}_n, \ldots, Y_n \equiv X_n - \bar{X}_n$. To this, consider the joint density function of the simple sample $X_1, \ldots, X_n$. Under the normality hypothesis for the random variable X, we have

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}\,\sigma^n_X}e^{-\frac{1}{2\sigma^2}\sum_{k=1}^{n}(x_k-\mu_X)^2}.$$

Hence, setting $\bar{x} \equiv \frac{1}{n}\sum_{k=1}^{n}x_k$, consider Then, the linear change of variables $\Psi : \mathbb{R}^n \to \mathbb{R}^n$ given by

$$\Psi(x_1, x_2, \ldots, x_n) \stackrel{\text{def}}{=} (\bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}).$$

The inverse $\Upsilon : \mathbb{R}^n \to \mathbb{R}^n$ of $\Psi : \mathbb{R}^n \to \mathbb{R}^n$ is given by

$$\Upsilon(y_1, y_2, \ldots, y_n) \stackrel{\text{def}}{=} (y_1 - \textstyle\sum_{k=2}^{n}y_k, y_1 + y_2, \ldots, y_1 + y_n).$$

In fact, we have

$$\begin{aligned} \Upsilon(\Psi(x_1, x_2, \ldots, x_n)) &= \Upsilon(\bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}) \\ &= (\bar{x} - \textstyle\sum_{k=2}^{n}(x_k - \bar{x}), \bar{x} + x_2 - \bar{x}, \ldots, \bar{x} + x_n - \bar{x}) \\ &= (x_1, x_2, \ldots, x_n), \end{aligned}$$

since,

$$\bar{x} - \sum_{k=2}^{n} (x_k - \bar{x}) = \bar{x} + (n-1)\,\bar{x} - \sum_{k=2}^{n} x_k = n\bar{x} + \sum_{k=2}^{n} x_k = \sum_{k=1}^{n} x_k + \sum_{k=2}^{n} x_k = x_1.$$

Moreover, setting

$$\bar{x} \equiv \frac{1}{n}\left(y_1 - \sum_{k=2}^{n} y_k + \sum_{k=2}^{n} (y_1 + y_k)\right) = \frac{1}{n}\left(y_1 + (n-1)\, y_1\right) = y_1$$

we have

$$\begin{aligned}
&\Psi\left(\Upsilon\left(y_1, y_2, \ldots, y_n\right)\right) = \Psi\left(y_1 - \sum_{k=2}^{n} y_k, y_1 + y_2, \ldots, y_1 + y_n\right) \\
&= \left(\bar{x}, y_1 + y_2 - \bar{x}, \ldots, y_1 + y_n - \bar{x}\right) = \left(y_1, y_1 + y_2 - y_1, \ldots, y_1 + y_n - y_1\right) \\
&= \left(y_1, y_2, \ldots, y_n\right).
\end{aligned}$$

Now, we have

$$\det\left(J_{\Upsilon}\left(y_1, y_2, \ldots, y_n\right)\right) = \begin{vmatrix} 1 & -1 & -1 & -1 & \cdots & -1 \\ 1 & 1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & 0 & \cdots & 1 \end{vmatrix} = n.$$

Therefore,

$$\begin{aligned}
f_{Y_1,\ldots,Y_n}\left(y_1, \ldots, y_n\right) &= f_{X_1,\ldots,X_n}\left(\Upsilon\left(y_1, y_2, \ldots, y_n\right)\right) \left|\det\left(J_{\Upsilon}\left(y_1, y_2, \ldots, y_n\right)\right)\right| \\
&= n f_{X_1,\ldots,X_n}\left(y_1 - \sum_{k=2}^{n} y_k, y_1 + y_2, \ldots, y_1 + y_n\right) \\
&= \frac{n}{(2\pi)^{n/2}\,\sigma^n} e^{-\frac{1}{2\sigma^2}\left(\left(y_1 - \sum_{k=2}^{n} y_k - \mu\right)^2 + \sum_{k=2}^{n} (y_1 + y_k - \mu)^2\right)} \\
&= \frac{n}{(2\pi)^{n/2}\,\sigma^n} e^{-\frac{1}{2\sigma^2}\left(n(y_1-\mu)^2 + \sum_{k=2}^{n} y_k^2 + \left(\sum_{k=2}^{n} y_k\right)^2\right)} \\
&= \frac{n^{1/2}}{(2\pi)^{1/2}\,\sigma} e^{-\frac{n}{2\sigma^2}(y_1-\mu)^2} \frac{n^{1/2}}{(2\pi)^{(n-1)/2}\,\sigma^{n-1}} e^{-\frac{1}{2\sigma^2}\left(\sum_{k=2}^{n} y_k^2 + \left(\sum_{k=2}^{n} y_k\right)^2\right)} \\
&= f_{Y_1}\left(y_1\right) g\left(y_2, \ldots, y_n\right),
\end{aligned}$$

because

$$\begin{aligned}
&\left(y_1 - \sum_{k=2}^{n} y_k - \mu\right)^2 + \sum_{k=2}^{n} \left(y_1 + y_k - \mu\right)^2 \\
&= y_1^2 + \left(\sum_{k=2}^{n} y_k\right)^2 + \mu^2 - 2y_1\left(\sum_{k=2}^{n} y_k\right) - 2\mu y_1 + 2\mu\left(\sum_{k=2}^{n} y_k\right) + \sum_{k=2}^{n} \left(y_1^2 + y_k^2 + \mu^2 + 2y_1 y_k - 2\mu y_1 - 2\mu y_k\right) \\
&= y_1^2 + \left(\sum_{k=2}^{n} y_k\right)^2 + \mu^2 - 2\mu y_1 + \sum_{k=2}^{n} y_1^2 + \sum_{k=2}^{n} y_k^2 + \sum_{k=2}^{n} \mu^2 - \sum_{k=2}^{n} 2\mu y_1 \\
&= n y_1^2 + n\mu^2 - 2n\mu y_1 + \left(\sum_{k=2}^{n} y_k\right)^2 + \sum_{k=2}^{n} y_k^2 \\
&= n\left(y_1 - \mu\right)^2 + \left(\sum_{k=2}^{n} y_k\right)^2 + \sum_{k=2}^{n} y_k^2.
\end{aligned}$$

This proves that the random variable $Y_1 \equiv \bar{X}_n$ is independent of $Y_2 \equiv X_2 - \bar{X}_n, \ldots, Y_n \equiv X_n - \bar{X}_n$. Hence, $\bar{X}_n$ is independent on $S^2_{X,n}$. □

Theorem 1069 *Assume that X is normally distributed with mean μ_X and variance σ_X^2. Then, the statistic $(n-1)\,S_{X,n}^2/\sigma_X^2$ has a chi-square distribution with $n-1$ degrees of freedom. In symbols,*

$$X \sim N\left(\mu_X, \sigma_X^2\right) \Rightarrow (n-1)\,S_{X,n}^2/\sigma_X^2 \sim \chi_{n-1}^2.$$

Proof. . □

Lemma 1070 *Let X and Y be real random variables on a probability space Ω such that X is standard Gaussian distributed, Y is chi-square distributed with n degrees of freedom, and X and Y are independent. Then, the statistic*

$$\frac{X}{\sqrt{Y/n}}$$

has the Student t-distribution with n degrees of freedom. In symbols,

$$X \sim N(0,1) \wedge Y \sim \chi_n^2 \wedge X \perp\!\!\!\perp Y \Rightarrow \frac{X}{\sqrt{Y/n}} \sim t_n.$$

Proof. . □

Definition 1071 *Consider the Borel function $g : \mathbb{R}^n \to \mathbb{R}$, given by*

$$g(x_1, \ldots, x_n) \stackrel{def}{=} \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(x_k - \frac{1}{n}\sum_{\ell=1}^{n} x_\ell\right)^2}, \quad \forall (x_1, \ldots, x_n) \in \mathbb{R}^n. \tag{13.42}$$

We call the statistic $S_n : \not\in \to \mathbb{R}$ given by

$$S_{X,n} \stackrel{def}{=} g \circ (X_1, \ldots, X_n) \equiv \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(X_k - \bar{X}_n\right)^2} \equiv \sqrt{S_{X,n}^2} \tag{13.43}$$

the sample standard deviation *of size n drawn from X.*

Theorem 1072 *Assume that X is Gaussian distributed with mean μ_X and variance σ_X^2. Then, the statistic*

$$\frac{\bar{X}_n - \mu_X}{S_n\sqrt{n}}$$

has the Student t-distribution with $n-1$ degrees of freedom. In symbols,

$$X \sim N\left(\mu_X, \sigma_X^2\right) \Rightarrow \frac{\bar{X}_n - \mu_X}{S_n\sqrt{n}} \sim t_{n-1}.$$

Proof. . □

Corollary 1073 *Assume that X is Gaussian distributed with mean μ_X and variance σ_X^2. Then, the statistic*

$$\frac{\bar{X}_n - \mu_X}{S_n\sqrt{n}}$$

is asymptotically standard Gaussian distribited. In symbols,

$$X \sim N\left(\mu_X, \sigma_X^2\right) \Rightarrow \frac{\bar{X}_n - \mu_X}{S_n\sqrt{n}} \stackrel{a}{\sim} N(0,1).$$

Theorem 1074 *Assume that X has finite moment of order 4. Then, the statistic*

$$\frac{\bar{X}_n - \mu_X}{S_n\sqrt{n}},$$

where $\mu_X \equiv \mathbf{E}[X]$, is asymptotically standard Gaussian distribited. In symbols,

$$\frac{\bar{X}_n - \mu_X}{S_n\sqrt{n}} \overset{a}{\sim} N(0,1)$$

Proof. . □

Let X [resp. Y] be a real random variable with finite mean μ_X [resp. μ_Y] and variance σ_X^2 [resp. σ_Y^2]. Let $X_1, \ldots, X_n$ [resp. $Y_1, \ldots, Y_n$] be a simple random sample of size n drawn from X, for some $n \in \mathbb{N}$, and let $\bar{X}_n$ [resp. $\bar{Y}_n$] and $S_{X,n}^2$ [resp. $S_{Y,n}^2$] be the sample mean and sample variance of size n drawn from X [resp. Y].

Definition 1075 *Consider the Borel function $g : \mathbb{R}^{2n} \to \mathbb{R}$, given by*

$$g(x_1, \ldots, x_n, y_1, \ldots, y_n) \overset{def}{=} \frac{1}{n-1}\sum_{k=1}^{n}\left(x_k - \frac{1}{n}\sum_{\ell=1}^{n} x_\ell\right)\left(y_k - \frac{1}{n}\sum_{\ell=1}^{n} y_\ell\right), \quad \forall (x_1, \ldots, x_n) \in \mathbb{R}^n. \tag{13.44}$$

We call the statistic $S_{X,Y,n} : \nexists \to \mathbb{R}$ given by

$$S_{X,Y,n} \overset{def}{=} g \circ (X_1, \ldots, X_n, Y_1, \ldots, Y_n) \equiv \frac{1}{n-1}\sum_{k=1}^{n}\left(X_k - \bar{X}_n\right)\left(Y_k - \bar{Y}_n\right) \tag{13.45}$$

the sample covariance *of size n drawn from X and Y.*

Proposition 1076 *We have*

$$S_{X,Y,n} = \frac{1}{n-1}\left(\sum_{k=1}^{n} X_k Y_k - n\bar{X}_n\bar{Y}_n\right). \tag{13.46}$$

Proof. A straightforward computation yields

$$\begin{aligned}
\sum_{k=1}^{n}\left(X_k - \bar{X}_n\right)\left(Y_k - \bar{Y}_n\right) &= \sum_{k=1}^{n}\left(X_k Y_k - X_k\bar{Y}_n - Y_k\bar{X}_n + \bar{X}_n\bar{Y}_n\right) \\
&= \sum_{k=1}^{n} X_k Y_k - \bar{Y}_n\sum_{k=1}^{n} X_k - \bar{X}_n\sum_{k=1}^{n} Y_k + \sum_{k=1}^{n}\bar{X}_n\bar{Y}_n \\
&= \sum_{k=1}^{n} X_n Y_n - n\bar{X}_n\bar{Y}_n - n\bar{X}_n\bar{Y}_n + n\bar{X}_n\bar{Y}_n \\
&= \sum_{k=1}^{n} X_n Y_n - n\bar{X}_n\bar{Y}_n,
\end{aligned}$$

from which, considering (**??**) of Definition 1075, the desired (**??**) immediately follows. □

Proposition 1077 *We have*

$$S_{X,Y,n} = \frac{1}{2n\left(n-1\right)} \sum_{j,k=1}^{n} \left(X_k - X_j\right)\left(Y_k - Y_j\right). \tag{13.47}$$

Proof. Thanks to a tricky manipulation, we can write

$$\begin{aligned}
\sum_{k=1}^{n} X_k Y_k - n\bar{X}_n\bar{Y}_n &= \sum_{k=1}^{n} X_k Y_k - n\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right)\left(\frac{1}{n}\sum_{k=1}^{n} Y_k\right) \\
&= \sum_{k=1}^{n} X_k Y_k - \frac{1}{n}\sum_{j,k=1}^{n} X_j Y_k \\
&= \frac{1}{2n}\sum_{j=1}^{n}\sum_{k=1}^{n} X_k Y_k + \frac{1}{2n}\sum_{k=1}^{n}\sum_{j=1}^{n} X_j Y_j - \frac{1}{2n}\sum_{j,k=1}^{n} X_j Y_k - \frac{1}{2n}\sum_{j,k=1}^{n} X_k Y_j \\
&= \frac{1}{2n}\sum_{j,k=1}^{n}\left(X_k Y_k - X_k Y_j - X_j Y_k + X_j Y_j\right) \\
&= \frac{1}{2n}\sum_{j,k=1}^{n}\left(X_k - X_j\right)\left(Y_k - Y_j\right),
\end{aligned}$$

which, considering (**??**) of Proposition 1076, implies (**??**). □

Definition 1078 *Assume that $S_{X,n}$ and $S_{Y,n}$ are non-zero. Then, we call the statistic*

$$R_{X,Y,n} \overset{def}{=} \frac{S_{X,Y,n}}{S_{X,n}S_{Y,n}} \tag{13.48}$$

the sample correlation *of size n drawn from X and Y.*

Definition 1079 *Given $k \in \mathbb{N}$, we call the statistic*

$$\bar{X}_n^k \overset{def}{=} \frac{1}{n}\sum_{j=1}^{n} X_j^k \tag{13.49}$$

the sample raw moment of order k *or kth* sample raw moment *drawn from X. Clearly,*

$$\bar{X}_n^1 = \bar{X}_n.$$

Proposition 1080 *Assume that X has finite moment of order k, for some $k \in \mathbb{N}$. Then,*

$$\mathbf{E}\left[\bar{X}_n^k\right] = \mu_k', \tag{13.50}$$

where $\mu_k' \equiv \mathbf{E}\left[X^k\right]$. In words, $\bar{X}_n^k$ is an unbiased estimator for the raw moment of order k of X.

Proof. A straightforward computation yields

$$\mathbf{E}\left[\bar{X}_n^k\right] = \mathbf{E}\left[\frac{1}{n}\sum_{j=1}^{n} X_j^k\right] = \frac{1}{n}\sum_{j=1}^{n}\mathbf{E}\left[X_j^k\right] = \frac{1}{n}\sum_{j=1}^{n}\mathbf{E}\left[X^k\right] = \mathbf{E}\left[X^k\right],$$

as desired. □

Proposition 1081 *Assume that X has finite moment of order k, for some $k \in \mathbb{N}$. Then,*

$$\bar{X}_n^k \xrightarrow{\boldsymbol{P}} \mu'_k,$$

where $\mu'_k \equiv \mathbf{E}\left[X^k\right]$. In words, $\bar{X}_n^k$ is an estimator for the raw moment of order k of X, which is consistent in probability.

Proof. The result follows by the application of the Khintchine Theorem 1008 to the sequence $\left(X_n^k\right)_{n\geq 1}$ of i.i.d. random variables with finite moment of order 1.

Proposition 1082 *Assume that X has finite moment of order $2k$, for some $k \in \mathbb{N}$. Then,*

$$\mathbf{D}^2\left[\bar{X}_n^k\right] = \frac{1}{n}\sigma^2_{X^k} \tag{13.51}$$

where $\sigma^2_{X^k} \equiv \mathbf{D}^2\left[X^k\right]$.

Proof. Note that, since X has finite moment of order $2k$ we have $\mathbf{E}\left[X^{2k}\right] < \infty$ and also $\mathbf{E}\left[\bar{X}_n^{2k}\right] < \infty$. As a consequence, $\mathbf{D}^2\left[X^k\right] \in \mathbb{R}$ and also $\mathbf{D}^2\left[\bar{X}_n^k\right] \in \mathbb{R}$. Then, a straightforward computation yields

$$\mathbf{D}^2\left[\bar{X}_n^k\right] = \mathbf{D}^2\left[\frac{1}{n}\sum_{j=1}^n X_j^k\right] = \frac{1}{n^2}\sum_{j=1}^n \mathbf{D}^2\left[X_j^k\right] = \frac{1}{n^2}\sum_{j=1}^n \mathbf{D}^2\left[X^k\right] = \frac{1}{n}\mathbf{D}^2\left[X^k\right],$$

as desired. □

Corollary 1083 *Assume that X has finite moment of order $2k$, for some $k \in \mathbb{N}$. Then,*

$$\bar{X}_n^k \xrightarrow{\boldsymbol{L}^2} \mu'_k,$$

where $\mu'_k \equiv \mathbf{E}\left[X^k\right]$. $\bar{X}_n^k$ is an estimator for the raw moment of order k of X, which is consistent in mean square.

Proof. Since $\bar{X}_n^k$ is an unbiased estimator for μ'_k, we can write

$$\mathbf{E}\left[\left(\bar{X}_n^k - \mu'_k\right)^2\right] = \mathbf{E}\left[\left(\bar{X}_n^k - \mathbf{E}\left[\bar{X}_n^k\right]\right)^2\right] = \mathbf{D}^2\left[\bar{X}_n^k\right] = \frac{1}{n}\mathbf{D}^2\left[X^k\right].$$

The claim immediately follows. □

Definition 1084 *Given $k \in \mathbb{N}$, we call the statistic*

$$M_{X,n}^k \stackrel{def}{=} \frac{1}{n}\sum_{j=1}^n \left(X_j - \bar{X}_n\right)^k \tag{13.52}$$

the sample central moment *of order k or kth* sample central moment *drawn from X. Clearly*

$$M_{X,n}^1 = 0.$$

Proposition 1085 *Assume that X has finite moment of order k, for some $k \in \mathbb{N}$. Then,*

$$M_{X,n}^k \xrightarrow{\boldsymbol{P}} \mu_k,$$

where $\mu_k \equiv \mathbf{E}\left[(X - \mathbf{E}[X])^k\right]$. In words, $M_{X,n}^k$ is an estimator for the central moment of order k of X, which is consistent in probability.

Proof. We can write

$$\frac{1}{n}\sum_{j=1}^{n}\left(X_j - \bar{X}_n\right)^k = \frac{1}{n}\sum_{j=1}^{n}\sum_{h=1}^{k}\binom{k}{h}X_j^h\bar{X}_n^{k-h} = \sum_{h=1}^{k}\binom{k}{h}\left(\frac{1}{n}\sum_{j=1}^{n}X_j^h\right)\bar{X}_n^{k-h} = \sum_{h=1}^{k}\binom{k}{h}\bar{X}_n^h\bar{X}_n^{k-h}. \tag{13.53}$$

Now, since X has finite moment of order k, thanks to Proposition 1081, we know that

$$\bar{X}_n^h \xrightarrow{\mathbf{P}} \mu_h', \qquad \bar{X}_n^{k-h} \xrightarrow{\mathbf{P}} \mu_{k-h}'.$$

By virtue of Theorem 945, it follows that

$$\sum_{h=1}^{k}\binom{k}{h}\bar{X}_n^h\bar{X}_n^{k-h} \xrightarrow{\mathbf{P}} \sum_{h=1}^{k}\binom{k}{h}\mu_h'\mu_{k-h}'. \tag{13.54}$$

On the other hand,

$$\mathbf{E}\left[(X - \mathbf{E}[X])^k\right] = \mathbf{E}\left[\sum_{h=1}^{k}\binom{k}{h}X^h\mathbf{E}\left[X^{k-h}\right]\right] = \sum_{h=1}^{k}\binom{k}{h}\mathbf{E}\left[X^h\mathbf{E}\left[X^{k-h}\right]\right] = \sum_{h=1}^{k}\binom{k}{h}\mathbf{E}\left[X^h\right]\mathbf{E}\left[X^{k-h}\right]. \tag{13.55}$$

Combining (13.54) and (13.55) we obtain the desired result.

Proposition 1086 *Assume that X has finite moment of order 2. Then, we have*

$$\mathbf{E}\left[M_{X,n}^2\right] = \frac{n-1}{n}\mu_2, \tag{13.56}$$

where $\mu_2 \equiv \mathbf{E}\left[(X - \mathbf{E}[X])^2\right]$.

Proof. . □

Proposition 1087 *Assume that X has finite moment of order 3. Then, we have*

$$\mathbf{E}\left[M_{X,n}^3\right] = \frac{(n-1)(n-2)}{n^2}\mu_3, \tag{13.57}$$

where $\mu_3 \equiv \mathbf{E}\left[(X - \mathbf{E}[X])^3\right]$.

Proof. . □

Proposition 1088 *Assume that X has finite moment of order 4. Then, we have*

$$\mathbf{E}\left[M_{X,n}^4\right] = \frac{(n-1)}{n^2}\left(\left(n^2 - 3n + 3\right)\mu_4 + 3\left(2n - 3\right)\mu_2^2\right), \tag{13.58}$$

where $\mu_k \equiv E\left[(X - E[X])^k\right]$, for $k = 2, 4$.

Proof. . □

Proposition 1089 *Assume that X has finite moment of order 4. Then, we have*

$$\mathbf{D}^2\left[M_{X,n}^2\right] = \frac{n-1}{n^3}\left((n-1)\,\mu_4 - (n-3)\,\mu_2^2\right),$$

where $\mu_k \equiv E\left[\left(X - E\left[X\right]\right)^k\right]$, for $k = 2, 4$.

Proof. . □

Corollary 1090 *Assume that X has finite moment of order 4. Then, the estimator $M_{X,n}^2$ for μ_2 is consistent in mean square.*

Proof. We can write

$$\begin{aligned}
\mathbf{E}\left[\left(M_{X,n}^2 - \mu_2\right)^2\right] &= \mathbf{E}\left[\left(M_{X,n}^2 - \frac{n-1}{n}\mu_2 + \frac{1}{n}\mu_2\right)^2\right] \\
&= \mathbf{E}\left[\left(M_{X,n}^2 - \mathbf{E}\left[M_{X,n}^2\right] + \frac{1}{n}\mu_2\right)^2\right] \\
&= \mathbf{E}\left[\left(M_{X,n}^2 - \mathbf{E}\left[M_{X,n}^2\right]\right)^2 + \left(M_{X,n}^2 - \mathbf{E}\left[M_{X,n}^2\right]\right)\frac{1}{n}\mu_2 + \frac{1}{n^2}\mu_2^2\right] \\
&= \mathbf{D}^2\left[M_{X,n}^2\right] + \frac{1}{n^2}\mu_2^2 \\
&= \frac{n-1}{n^3}\left((n-1)\,\mu_4 - (n-3)\,\mu_2^2\right) + \frac{1}{n^2}\mu_2^2
\end{aligned}$$

The claim immediately follows. □

Definition 1091 *We call the statistic*

$$Skew_{X,n}^{(g_1)} \overset{def}{=} \frac{M_{X,n}^3}{\left(M_{X,n}^2\right)^{3/2}} \equiv \frac{\frac{1}{n}\sum_{k=1}^n \left(X_k - \bar{X}_n\right)^3}{\left[\frac{1}{n}\sum_{k=1}^n \left(X_k - \bar{X}_n\right)^2\right]^{3/2}} \tag{13.59}$$

the g_1-sample skewness *of size n drawn from X. We call the statistic*

$$Skew_{X,n}^{(b_1)} \overset{def}{=} \frac{M_{X,n}^3}{\left(S_{X,n}^2\right)^{3/2}} \equiv \frac{\frac{1}{n}\sum_{k=1}^n \left(X_k - \bar{X}_n\right)^3}{\left[\frac{1}{n-1}\sum_{k=1}^n \left(X_k - \bar{X}_n\right)^2\right]^{3/2}} \tag{13.60}$$

the b_1-sample skewness *of size n drawn from X.*

Remark 1092 *We have*

$$Skew_{X,n}^{(b_1)} = \frac{(n-1)^{3/2}}{n^{3/2}} Skew_{X,n}^{(g_1)}. \tag{13.61}$$

Definition 1093 *We call the statistic.*

$$Skew_{X,n}^{(G_1)} \overset{def}{=} \frac{\sqrt{n\,(n-1)}}{n-2} Skew_{X,n}^{(g_1)} \tag{13.62}$$

the G_1-sample skewness *of size n drawn from X.*

Remark 1094 *We have*

$$Skew_{X,n}^{(G_1)} = \frac{n^2}{(n-1)(n-2)} Skew_{X,n}^{(b_1)}.$$

Proposition 1095 *Assume that X has finite moment of order 3. We have*

$$\mathbf{E}\left[Skew_{X,n}^{(g_1)}\right] = \gamma_1 + o\left(\frac{1}{n}\right) \tag{13.63}$$

where $\gamma_1 \equiv \mu_3/\sigma^3 \equiv Skew(X)$, for $\mu_3 \equiv E\left[(X - E[X])^3\right]$ and $\sigma \equiv \mathbf{D}[X]$ (see [?, 27.7.8 - p. 357]).

Proposition 1096 *Assume that X has finite moment of order 6. We have*

$$\mathbf{D}^2\left[Skew_{X,n}^{(g_1)}\right] = \frac{4\mu_2^2\mu_6 - 12\mu_2\mu_3\mu_5 - 24\mu_2^3\mu_4 + 9\mu_3^2\mu_4 + 35\mu_2^2\mu_3^2 + 36\mu_2^5}{4n\mu_2^5} + o\left(\frac{1}{n^{3/3}}\right) \tag{13.64}$$

where $\mu_k \equiv E\left[(X - E[X])^3\right]$, for $k = 2, \ldots, 6$ (see [?, 27.7.8 - p. 357]).

Proposition 1097 *Assume that X is Gaussian distributed. We have*

$$\mathbf{E}\left[Skew_{X,n}^{(g_1)}\right] = 0 \quad and \quad \mathbf{D}^2\left[Skew_{X,n}^{(g_1)}\right] = \frac{6(n-2)}{(n+1)(n+3)} \tag{13.65}$$

Proof. See [?, 29.3.7, p. 386]. □

Corollary 1098 *Assume that X is Gaussian distributed. We have*

$$\mathbf{E}\left[Skew_{X,n}^{(b_1)}\right] = 0, \qquad \mathbf{D}^2\left[Skew_{X,n}^{(b_1)}\right] = \frac{6(n-2)(n-1)^3}{n^3(n+1)(n+3)}, \tag{13.66}$$

and

$$\mathbf{E}\left[Skew_{X,n}^{(G_1)}\right] = 0, \qquad \mathbf{D}^2\left[Skew_{X,n}^{(G_1)}\right] = \frac{6n(n-1)}{(n-2)(n+1)(n+3)}. \tag{13.67}$$

Proof. Equation 13.66 clearly follows from Equations 13.61 and 13.65... See [?, 29.3.9 - p. 387]. □

Definition 1099 *We call the statistic*

$$Kurt_{X,n}^{(g_2)} \stackrel{def}{=} \frac{M_{X,n}^4}{\left(M_{X,n}^2\right)^2} \equiv \frac{\frac{1}{n}\sum_{k=1}^n \left(X_k - \bar{X}_n\right)^3}{\left[\frac{1}{n}\sum_{k=1}^n \left(X_k - \bar{X}_n\right)^2\right]^2} \tag{13.68}$$

the g_2-sample kurtosis *of size n drawn from X. We call the statistic*

$$Kurt_{X,n}^{(b_2)} \stackrel{def}{=} \frac{M_{X,n}^2}{\left(S_{X,n}^2\right)^2} \equiv \frac{\frac{1}{n}\sum_{k=1}^n \left(X_k - \bar{X}_n\right)^3}{\left[\frac{1}{n-1}\sum_{k=1}^n \left(X_k - \bar{X}_n\right)^2\right]^2} \tag{13.69}$$

the b_2-sample kurtosis *of size n drawn from X.*

Remark 1100 *We have*

$$Kurt_{X,n}^{(b_2)} = \frac{(n-1)^2}{n^2} Kurt_{X,n}^{(g_2)}. \tag{13.70}$$

Definition 1101 *We call the statistic.*

$$Kurt_{X,n}^{(G_2)} \stackrel{def}{=} \frac{(n-1)}{(n-2)(n-3)} \left((n+1) Kurt_{X,n}^{(g_2)} + 6 \right) \tag{13.71}$$

the G_1-sample kurtosis *of size* n *drawn from* X*.*

Definition 1102 *We call the statistic*

$$Kurtex_{X,n}^{(g_2)} \stackrel{def}{=} Kurt_{X,n}^{(g_2)} - 3 \qquad [resp.\ Kurtex_{X,n}^{(b_2)} \stackrel{def}{=} Kurt_{X,n}^{(b_2)} - 3] \tag{13.72}$$

the g_2 *[resp.* b_2*]* -sample excess kurtosis *of size* n *drawn from* X*.*

Proposition 1103 *Assume that* X *has finite moment of order* 4*. We have*

$$\mathbf{E}\left[Kurtex_{X,n}^{(g_1)} \right] = \gamma_2 + o\left(\frac{1}{n} \right) \tag{13.73}$$

where $\gamma_2 \equiv \mu_4/\sigma^4 - 3 \equiv Kurtex(X)$*, for* $\mu_4 \equiv E\left[(X - E[X])^4 \right]$ *and* $\sigma \equiv \mathbf{D}[X]$ *(see [?, 27.7.8 - p. 357]).*

Proposition 1104 *Assume that* X *has finite moment of order* 8*. We have*

$$\mathbf{D}^2\left[Kurtex_{X,n}^{(g_1)} \right] = \frac{\mu_2^2\mu_8 - 4\mu_2\mu_4\mu_6 - 8\mu_2^2\mu_3\mu_5 + 4\mu_4^3 - \mu_2^2\mu_4^2 + 16\mu_2\mu_3^2\mu_4 + 16\mu_2^3\mu_3^2}{n\mu_2^6} + o\left(\frac{1}{n^{3/3}} \right) \tag{13.74}$$

where $\mu_k \equiv E\left[(X - E[X])^3 \right]$*, for* $k = 2, \ldots, 8$ *(see [?, 27.7.8 - p. 357]).*

Proposition 1105 *Assume that* X *is Gaussian distributed. We have*

$$\mathbf{E}\left[Kurtex_{X,n}^{(g_2)} \right] = -\frac{6}{n+1} \quad and \quad \mathbf{D}^2\left[Kurtex_{X,n}^{(g_2)} \right] = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)} \tag{13.75}$$

Proof. See [?, 29.3.7, p. 386]. □

Corollary 1106 *Assume that* X *is Gaussian distributed. We have*

$$\mathbf{E}\left[Kurtex_{X,n}^{(b_2)} \right] = 3\frac{(n-1)^3}{n^2(n+1)} - 3, \qquad \mathbf{D}^2\left[Kurtex_{X,n}^{(b_2)} \right] = \frac{24(n-1)^4(n-2)(n-3)}{n^3(n+1)^2(n+3)(n+5)}, \tag{13.76}$$

and

$$\mathbf{E}\left[Kurt_{X,n}^{(G_2)} \right] = 0, \qquad \mathbf{D}^2\left[Kurt_{X,n}^{(G_2)} \right] = \frac{24n(n-1)^2}{(n-3)(n-2)(n+3)(n+5)}. \tag{13.77}$$

Proof. Equation 13.76 clearly follows from Equations 13.70, 13.72, and 13.75... See [?, 29.3.9 - p. 387]. □

Chapter 14

Point Estimation

Let X be a real random variable a probability space $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ representing a population with distribution depending on an (unknown) parameter vector $\theta \in \Theta \subseteq \mathbb{R}^M$, for some $M \in \mathbb{N}$. We will write $P_X(\cdot\,;\theta) : \mathcal{B}(\mathbb{R}) \times \Theta \to \mathbb{R}$, more briefly $P_X(\cdot\,;\theta)$ [resp. $F_X(\cdot\,;\theta) : \mathbb{R} \times \Theta \to \mathbb{R}$, more briefly $F_X(\cdot\,;\theta)$], for the distribution [resp. the distribution function] of X to enhance their dependence on θ. Similarly, under the assumption that X is absolutely continuous, we will write $f_X(\cdot\,;\theta) : \mathbb{R} \times \Theta \to \mathbb{R}$, more briefly $f_X(\cdot\,;\theta)$, for the density function of X.

Example 1107 *Assume that the daily returns of a risky stock in a quarter are normally distributed with unknown mean $\mu \in \mathbb{R}$ and standard deviation $\sigma > 0$. We set $\theta \equiv (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+ \equiv \Theta$ and we can write*

$$P_X(B;\theta) \equiv P_X(B;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\int_B e^{-\frac{(x-\mu)^2}{2\sigma^2}} d\mu(x), \quad \forall B \in \mathcal{B}(\mathbb{R}),\ (\mu,\sigma) \in \mathbb{R}\times\mathbb{R}_+$$

$$F_X(x;\theta) \equiv F_X(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x} e^{-\frac{(u-\mu)^2}{2\sigma^2}} du, \quad \forall x \in \mathbb{R},\ (\mu,\sigma) \in \mathbb{R}\times\mathbb{R}_+,$$

$$f_X(x;\theta) \equiv f_X(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(u-\mu)^2}{2\sigma^2}}, \quad \forall x \in \mathbb{R},\ (\mu,\sigma) \in \mathbb{R}\times\mathbb{R}_+.$$

Definition 1108 *We call the parameters μ and σ introduced in Example 1107 the quarterly* mean *and* standard deviation *of the stock return (population).*

Let $X_1, \ldots, X_n$ be a simple random sample of size n drawn from X.

Notation 1109 *When X has distribution $P_X(\cdot\,;\theta) : \mathcal{B}(\mathbb{R}) \times \Theta \to \mathbb{R}$ [resp. distribution function $F_X(\cdot\,;\theta) : \mathbb{R} \times \Theta \to \mathbb{R}$], where θ is a parameter vector in $\Theta \subseteq \mathbb{R}^M$. We will denote by $P_{X_1,\ldots,X_n}(\cdot\,;\theta) : \mathcal{B}(\mathbb{R}^n) \times \Theta \to \mathbb{R}$, more briefly $P_{X_1,\ldots,X_n}(\cdot\,;\theta)$ [resp. $F_{X_1,\ldots,X_n}(\cdot\,;\theta) : \mathbb{R}^n \times \Theta \to \mathbb{R}$, more briefly $F_{X_1,\ldots,X_n}(\cdot\,;\theta)$], the joint distribution [resp distribution function] of the sample $X_1, \ldots, X_n$, which is given by*

$$P_{X_1,\ldots,X_n}(\mathsf{X}_{k=1}^n B_k;\theta) = \prod_{k=1}^n P_{X_k}(B_k;\theta) = \prod_{k=1}^n P_X(B_k;\theta), \quad \forall B_1,\ldots,B_n \in \mathcal{B}(\mathbb{R}),\ \theta \in \Theta,$$

[resp. $F_{X_1,\ldots,X_n}(x_1,\ldots,x_n;\theta) = \prod_{k=1}^n F_{X_k}(x_k;\theta) = \prod_{k=1}^n F_X(x_k;\theta), \quad \forall (x_1,\ldots,x_n) \in \mathbb{R}^n,\ \theta \in \Theta$].

When X is absolutely continuous with density function $f_X(\cdot\,;\theta) : \mathbb{R} \times \Theta \to \mathbb{R}_+$, we will denote by $f_{X_1,\ldots,X_n}(\cdot\,;\theta) : \mathbb{R}^n \to \mathbb{R}_+$, more briefly $f_{X_1,\ldots,X_n}(\cdot\,;\theta)$, the joint density function of the sample $X_1, \ldots, X_n$, which is given by

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n;\theta) = \prod_{k=1}^n f_{X_k}(x_k;\theta) = \prod_{k=1}^n f_X(x_k;\theta), \quad \forall (x_1,\ldots,x_n) \in \mathbb{R}^n,\ \theta \in \Theta.$$

The goal of a point estimation is to use a simple random sample $X_1, \ldots, X_n$ of suitable size $n \in \mathbb{N}$ drawn from X and a Borel map $g : \mathbb{R}^n \to \mathbb{R}^M$ such that by means of the statistic $G_n : \Omega \to \mathbb{R}^M$ given by

$$G_n \overset{\text{def}}{=} g \circ (X_1, \ldots, X_n)$$

(see (1028)) we can make a "good guess" for the true value of θ.

Definition 1110 *A statistic $G_n : \Omega \to \mathbb{R}^M$ given by (**??**) is called a* point estimator *of the parameter θ if we can exploit the realizations $G_n(\omega)$ of G_n to estimate the true value of θ. Given any realization $x_1, \ldots, x_n$ of the sample $X_1, \ldots, X_n$, where $x_k = X_k(\omega)$, for some $\omega \in \Omega$ and every $k = 1, \ldots, n$, the real number*

$$G_n(\omega) \overset{def}{=} g(X_1(\omega), \ldots, X_n(\omega)) \equiv g(x_1, \ldots, x_n) \tag{14.1}$$

is called a point estimate *of θ.*

Example 1111 *With reference to Example 1107, consider the Borel map $g : \mathbb{R}^n \to \mathbb{R}^2$ given by*

$$g(x_1, \ldots, x_n) \overset{def}{=} \left(\frac{1}{n}\sum_{k=1}^n x_k, \frac{1}{n-1}\sum_{k=1}^n \left(x_k - \frac{1}{n}\sum_{\ell=1}^n x_\ell \right) \right), \quad \forall\, (x_1, \ldots, x_n) \in \mathbb{R}^n.$$

Then, the statistic $G_n : \Omega \to \mathbb{R}^2$ given by

$$G_n(\omega) \overset{def}{=} \left(\frac{1}{n}\sum_{k=1}^n X_k(\omega), \frac{1}{n-1}\sum_{k=1}^n \left(X_k(\omega) - \frac{1}{n}\sum_{\ell=1}^n X_\ell(\omega) \right) \right), \quad \forall \omega \in \Omega,$$

equivalently written as

$$G_n \overset{def}{=} \left(\bar{X}_n, S^2_{X,n} \right),$$

is a point estimator for the quarterly mean μ and variance σ^2 of the stock returns.

Discussion. In fact, from (**??**) [resp. (**??**)] of Proposition **??** [resp. 1064], we know that

$$\mathbf{E}\left[\bar{X}_n\right] = \mu \quad [\text{resp. } \mathbf{E}\left[S^2_{X,n}\right] = \sigma^2].$$

Hence,

$$\mathbf{E}[G_n] = \left(\mathbf{E}\left[\bar{X}_n\right], \mathbf{E}\left[S^2_{X,n}\right] \right) = \left(\mu, \sigma^2\right).$$

□

Notation 1112 *As it is customary, we will use the symbol $\hat{\theta}_n$ to denote a point estimator of θ derived from a simple random sample of size n. In terms of the above notation $\hat{\theta}_n \equiv G_n$. We will also use the symbol $\hat{\theta}$ when the reference to the size n of the sample is not particularly relevant. We will use the notation $\hat{\theta}_n(\omega)$ [resp. $\hat{\theta}(\omega)$] to denote the sample point estimate of θ, given by the value taken by the estimator $\hat{\theta}_n$ [resp. $\hat{\theta}$] on the occurrence of the sample point ω. For instance, given a random variable X of unknown mean μ, the notation $\hat{\mu}_n \equiv \bar{X}_n$ will mean that we are using the sample mean $\bar{X}_n$ as a point estimator of μ and the notation $\hat{\mu}_n(\omega) \equiv \bar{X}_n(\omega) \equiv \frac{1}{n}\sum_{k=1}^n x_k$, for the realization $x_1 \equiv X_1(\omega), \ldots, x_n \equiv X_n(\omega)$ of a simple random sample $X_1, \ldots, X_n$ of size n drawn from X on the occurrence of a sample point ω, will mean that $\hat{\mu}_n(\omega)$ is the sample estimate of μ which results on the occurrence of the sample point ω.*

Example 1113 *With reference to Examples 1107 and 1111, consider the following sequence of (adjusted) daily closing prices (in Euros) of STMicroelectronics stock in Italian FTSE-MIB from* 2017.01.16 *to* 2017.01.31

01.16	01.17	01.18	01.19	01.20	01.23	01.24	01.25	01.26	01.27	01.31	01.31
10.73	10.66	10.73	10.84	10.80	10.73	10.99	11.13	12.03	12.47	12.42	12.12

With reference to the sample mean [resp. sample variance] estimator

$$\hat{\mu}_{12} \equiv \frac{1}{12}\sum_{k=1}^{12} P_k, \qquad [\text{resp. } \hat{\sigma}^2_{12} \equiv \frac{1}{11}\sum_{k=1}^{12}\left(P_k - \frac{1}{12}\sum_{k=1}^{12} P_k\right)^2],$$

usually called the mean price *or* expected price *[resp.* price variance*],we have the sample estimate*

$$\hat{\mu}_{12}(\omega) \equiv \frac{1}{12}\sum_{k=1}^{12} p_k = 11.30, \qquad [\text{resp. } \hat{\sigma}^2_{12}(\omega) \equiv \frac{1}{11}\sum_{k=1}^{12}\left(p_k - \frac{1}{12}\sum_{k=1}^{12} p_k\right)^2 = 0.53].$$

Note that, in this context, ω represents the sample point in the space of all possible paths of the price process $(P_t)_{t\geq 0}$ such that $P_{01.16}(\omega) = 10.73$, $P_{01.17}(\omega) = 10.66$, ..., $P_{01.31}(\omega) = 12.12$. In terms of the returns of the stock, which are defined as

$$R_{k+1} = \frac{P_{k+1} - P_k}{P_k}, \quad \forall k = 1, \ldots, 11$$

we have

01.17	01.18	01.19	01.20	01.23	01.24	01.25	01.26	01.27	01.31	01.31
−0.0065	0.0066	0.0103	−0.0037	−0.0065	0.0242	0.0127	0.0809	0.0366	−0.0040	−0.0242

Therefore, with reference to the sample mean [resp. sample variance] estimator

$$\hat{\mu}_{11} \equiv \frac{1}{11}\sum_{k=1}^{11} R_{k+1}, \qquad [\text{resp. } \hat{\sigma}^2_{12} \equiv \frac{1}{10}\sum_{k=1}^{11}\left(R_{k+1} - \frac{1}{11}\sum_{k=1}^{11} R_{k+1}\right)^2],$$

usually called the mean return *or* expected return *[resp.* return variance*], we have the sample estimate*

$$\hat{\mu}_{11}(\omega) \equiv \frac{1}{11}\sum_{k=1}^{11} r_{k+1} = 0.0115, \qquad [\text{resp. } \hat{\sigma}^2_{11}(\omega) \equiv \frac{1}{10}\sum_{k=1}^{11}\left(r_k - \frac{1}{11}\sum_{k=1}^{11} r_k\right)^2 = 0.0008].$$

Definition 1114 *We call the* mean square error *of a point estimator $\hat{\theta}$ in the estimation of the true value of a parameter θ the positive number*

$$\mathbf{MSE}\left(\hat{\theta}\right) \overset{def}{=} \mathbf{E}\left[\left(\hat{\theta} - \theta\right)^2\right].$$

The better is an estimator the smaller is its mean square error for all values of the parameter $\theta \in \Theta$.

Seeking an estimator whose mean square error is the minimum possible for all values of the parameter θ is generally a too ambitious goal. The standard procedure is to restict the set of estimators under consideration according to some criterion and seek for the estimetors which is the best in the restricted set.

Definition 1115 *We call the* bias *of a point estimator $\hat{\theta}$ in the estimation of the true value of a parameter θ the real number*

$$\mathbf{Bias}\left(\hat{\theta}\right) \overset{def}{=} \mathbf{E}\left[\hat{\theta}\right] - \theta.$$

Definition 1116 *A point estimator $\hat{\theta}$ of a parameter θ is said to be* unbiased *if*

$$\mathbf{Bias}\left(\hat{\theta}\right) = 0, \quad \forall \theta \in \Theta.$$

Example 1117 *Given a random variable X with finite mean μ and variance σ^2, for any $n \in \mathbb{N}$, the sample mean $\bar{X}_n$ and the unbiased sample variance $S^2_{X,n}$ drawn from X are unbiased point estimators of μ and σ^2, respectively. On the contrary, the biased sample variance $\tilde{S}^2_{X,n}$ is a biased estimator of σ^2 and we have*

$$\mathbf{Bias}\left(\tilde{S}^2_{X,n}\right) = \mathbf{E}\left[\tilde{S}^2_{X,n}\right] - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{1}{n}\sigma^2.$$

Discussion. See Propositions **??**, 1064, and 1067. □

Proposition 1118 *We have*

$$\mathbf{MSE}\left(\hat{\theta}\right) = \mathbf{D}^2\left[\hat{\theta}\right] + \mathbf{Bias}\left(\hat{\theta}\right)^2.$$

Proof. We can clearly write

$$\begin{aligned}
\mathbf{MSE}\left(\hat{\theta}_n\right) &= \mathbf{E}\left[\left(\hat{\theta} - \theta\right)^2\right] \\
&= \mathbf{E}\left[\left(\hat{\theta} - \mathbf{E}\left[\hat{\theta}\right] + \mathbf{E}\left[\hat{\theta}\right] - \theta\right)^2\right] \\
&= \mathbf{E}\left[\left(\hat{\theta} - \mathbf{E}\left[\hat{\theta}\right]\right)^2 + 2\left(\hat{\theta} - \mathbf{E}\left[\hat{\theta}\right]\right)\left(\mathbf{E}\left[\hat{\theta}\right] - \theta\right) + \left(\mathbf{E}\left[\hat{\theta}\right] - \theta\right)^2\right] \\
&= \mathbf{E}\left[\left(\hat{\theta} - \mathbf{E}\left[\hat{\theta}\right]\right)^2\right] + 2\mathbf{E}\left[\left(\hat{\theta} - \mathbf{E}\left[\hat{\theta}\right]\right)\left(\mathbf{E}\left[\hat{\theta}\right] - \theta\right)\right] + \mathbf{E}\left[\left(\mathbf{E}\left[\hat{\theta}\right] - \theta\right)^2\right].
\end{aligned}$$

On the other hand, since θ and $\mathbf{E}\left[\hat{\theta}\right]$ are real numbers

$$\mathbf{E}\left[\left(\hat{\theta} - \mathbf{E}\left[\hat{\theta}\right]\right)\left(\mathbf{E}\left[\hat{\theta}\right] - \theta\right)\right] = \left(\mathbf{E}\left[\hat{\theta}\right] - \theta\right)\mathbf{E}\left[\hat{\theta} - \mathbf{E}\left[\hat{\theta}\right]\right] = \left(\mathbf{E}\left[\hat{\theta}\right] - \theta\right)\left(\mathbf{E}\left[\hat{\theta}\right] - \mathbf{E}\left[\hat{\theta}\right]\right) = 0$$

and

$$\mathbf{E}\left[\left(\mathbf{E}\left[\hat{\theta}\right] - \theta\right)^2\right] = \left(\mathbf{E}\left[\hat{\theta}\right] - \theta\right)^2.$$

The desired claim immediately follows. □

Definition 1119 *We call the* standard error *of a point estimator $\hat{\theta}$ (in the estimation of the true value of a parameter θ) the positive number*

$$\mathbf{SE}\left(\hat{\theta}\right) \overset{def}{=} \sqrt{\mathbf{D}^2\left[\hat{\theta}\right]}. \tag{14.2}$$

Definition 1120 *We say that an estimator $\hat{\theta}_n$ of θ is (*asymptotically*)* consistent in probability *[resp.* mean square*] if*

$$\hat{\theta}_n \xrightarrow{\mathbf{P}} \theta \qquad [\text{resp. } \hat{\theta}_n \xrightarrow{\mathbf{L}^2} \theta] \tag{14.3}$$

as $n \to \infty$.

Proposition 1121 *An unbiased estimator $\hat{\theta}_n$ of θ is consistent in mean square if and only if*

$$\lim_{n\to\infty} \mathbf{D}^2\left[\hat{\theta}_n\right] = 0. \tag{14.4}$$

Proof. Since the estimator is unbiased, we can write

$$\mathbf{E}\left[\left(\hat{\theta}_n - \theta\right)^2\right] = \mathbf{E}\left[\left(\hat{\theta}_n - \mathbf{E}\left[\hat{\theta}_n\right]\right)^2\right] = \mathbf{D}^2\left[\hat{\theta}_n\right].$$

The claim immediately follows. □

Proposition 1122 *If an unbiased estimator $\hat{\theta}_n$ of θ is consistent in mean square Then, it is also consistent in probability.*

Proof. Since the estimator is unbiased, by applying the Chebyshev inequality, we can write

$$\mathbf{P}\left(\left|\hat{\theta}_n - \theta\right| \geq \varepsilon\right) = \mathbf{P}\left(\left|\hat{\theta}_n - \mathbf{E}\left[\hat{\theta}_n\right]\right| \geq \varepsilon\right) \leq \frac{1}{\varepsilon^2}\mathbf{E}\left[\left(\hat{\theta}_n - \mathbf{E}\left[\hat{\theta}_n\right]\right)^2\right] = \frac{1}{\varepsilon^2}\mathbf{E}\left[\left(\hat{\theta}_n - \theta\right)^2\right],$$

for every $n \in \mathbb{N}$ and every $\varepsilon \in \mathbb{R}_+$. Therefore, from the convergence of $\hat{\theta}_n$ to θ in mean square it follows the convergence of $\hat{\theta}_n$ to θ in probability. □

Example 1123 *Given a random variable X with finite mean μ and variance σ^2, for any $n \in \mathbb{N}$, the sample mean $\bar{X}_n$ is an estimator of μ which is consistent in mean square.*

Discussion. Since $\bar{X}_n$ is an unbiased estimator of μ, considering Equation (**??**), we can write

$$\mathbf{E}\left[\left(\bar{X}_n - \mu\right)^2\right] = \mathbf{E}\left[\left(\bar{X}_n - \mathbf{E}\left[\bar{X}_n\right]\right)^2\right] = \mathbf{D}^2\left[\bar{X}_n\right] = \frac{\sigma^2}{n},$$

for every $n \in \mathbb{N}$. Passing to the limit as n goes to infinity, the claim follows. □

14.1 Methods of Moments

The idea of the methods of moment is that some unknown parameter appearing in the distribution of a real random variable can be etimated by equating a suitable number of moments of the random variable, with progressively increasing order, to the corresponding *sample moments*.

Let X be a real random variable on a probability space $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ representing a population with distribution $P_X(\cdot\,;\theta) : \mathcal{B}(\mathbb{R}) \times \Theta \to \mathbb{R}$ depending on a parameter $\theta \in \Theta \subseteq \mathbb{R}^M$, for some $M \in \mathbb{N}$, and let $X_1, \ldots, X_n$ be a simple random sample of size n, for some $n \in \mathbb{N}$, drawn from X.

Definition 1124 *For any $M \in \mathbb{N}$ such that $\mathbf{E}\left[|X|^M\right] < \infty$, we call the Mth* population raw moment *the Mth moment of the random variable X, that is the function $\mu'_1 : \Theta \to \mathbb{R}$ of the parameter θ given by*

$$\mu'_1(\theta) \stackrel{def}{=} \mathbf{E}\left[X^M\right] = \int_{\mathbb{R}} x^M dP_X(x;\theta), \quad \forall \theta \in \Theta. \tag{14.5}$$

We call the Mth sample moment *the statistic*

$$\bar{X}_n^{(M)} \stackrel{def}{=} \frac{1}{n}\sum_{k=1}^n X_k^M. \tag{14.6}$$

Example 1125 *The first population raw moment $\mu'_1 : \Theta \to \mathbb{R}$ is given by*

$$\mu'_1(\theta) \stackrel{def}{=} \mathbf{E}[X] = \int_{\mathbb{R}} x dP_X(x;\theta), \quad \forall \theta \in \Theta;$$

the first sample moment is the sample mean

$$\bar{X}_n^{(1)} \stackrel{def}{=} \frac{1}{n}\sum_{k=1}^n X_k \equiv \bar{X}_n.$$

The second population raw moment $\mu'_2 : \Theta \to \mathbb{R}$ is given by

$$\mu'_2(\theta) \stackrel{def}{=} \mathbf{E}\left[X^2\right] = \int_{\mathbb{R}} x^2 dP_X(x;\theta), \quad \forall \theta \in \Theta;$$

the second sample moment is given by

$$\bar{X}_n^{(2)} \stackrel{def}{=} \frac{1}{n}\sum_{k=1}^n X_k^2.$$

Remark 1126 *From (13.24) and (13.40), setting $\mu = 0$, it clearly follows*

$$\bar{X}_n^{(2)} = \tilde{S}_{X,n}^2 + \bar{X}_n^2, \tag{14.7}$$

where $\tilde{S}_{X,n}^2$ is the biased sample variance and $\bar{X}_n^2$ is the square of the sample mean $\bar{X}_n$.

Recall 1127 *If $\mathbf{E}\left[|X|^M\right] < \infty$ for some $M \in \mathbb{N}$, Then, $\mathbf{E}\left[|X|^m\right] < \infty$ for every $m = 1, \ldots, M$.*

Definition 1128 *The* method of moments *consists in determining the estimators $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(M)}$ of the entries $\theta_1, \ldots, \theta_M$ of the parameter vector θ by means of the following procedure:*

1. *replace the parameters $\theta_1, \ldots, \theta_M$, which appear in the population moments $\mu'_1(\theta), \ldots, \mu'_M(\theta)$ with the corresponding estimators $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(M)}$;*
2. *equate the first M population moments $\bar{X}_n, \ldots, \bar{X}_n^{(M)}$ to the corresponding modified M sample moments;*
3. *solve the resulting equations for $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(M)}$.*

Example 1129 *Let X be an exponential random variable on a probability space $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ with rate parameter $\lambda > 0$ (see Definition 638). In symbols, $X \tilde{} Exp(\lambda)$. A simple random sample $X_1, \ldots, X_n$ of size n, for some $n \in \mathbb{N}$, drawn from X, ha density $f_{X_1,\ldots,X_n} : \mathbb{R}^n \times \mathbb{R}_{++} \to \mathbb{R}$ given by*

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n; \lambda) = \prod_{k=1}^{n} \lambda e^{-\lambda x_k} 1_{\mathbb{R}_+}(x_k) = \lambda^n e^{-\lambda \sum_{k=1}^{n} x_k} 1_{\mathbb{R}_+^n}(x_1, \ldots x_n),$$

for every $(x_1, \ldots, x_n; \lambda) \in \mathbb{R}_+^n \times \mathbb{R}_{++}$. To build an estimator $\hat{\lambda}_n^M$ for λ by the method of moments we compute the first population raw moment. We have

$$\mu_1'(\lambda) = \mathbf{E}[X] = \frac{1}{\lambda}.$$

Thereafter, we replace $\hat{\lambda}_n^M$ to λ and equate the first population moment to the first sample moment. This yields

$$\frac{1}{\hat{\lambda}_n^M} = \bar{X}_n.$$

As a consequence, since we clearly have

$$\mathbf{P}\left(\bar{X}_n > 0\right) = 1,$$

solving for $\hat{\lambda}_n$ we obtain

$$\hat{\lambda}_n^M = \frac{1}{\bar{X}_n}.$$

We know that

$$\mathbf{E}\left[\bar{X}_n\right] = \mathbf{E}[X] = \frac{1}{\lambda} \quad \textit{and} \quad \mathbf{D}^2\left[\bar{X}_n\right] = \frac{1}{n}\mathbf{D}^2[X] = \frac{1}{n\lambda^2}$$

Hence,

$$\frac{1}{\mathbf{E}\left[\bar{X}_n\right]} = \lambda.$$

On the other hand, in general,

$$\frac{1}{\mathbf{E}\left[\bar{X}_n\right]} \neq \mathbf{E}\left[\frac{1}{\bar{X}_n}\right].$$

Therefore, it seems rather counterintuitive that $\hat{\lambda}_n$ is unbiases. However, we apply the so called delta method. Considering the function $f : (0,1) \to \mathbb{R}$ given by

$$f(x) \overset{def}{=} \frac{1}{x}, \quad \forall x \in (0,1)$$

by virtue of the Taylor formula, fixed any $x_0 \in (0,1)$, we can write

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0),$$

for every $x \in (0,1)$, where

$$f'(x_0) = -\frac{1}{x_0^2}.$$

Now, we have

$$\hat{\lambda}_n^M = \frac{1}{\bar{X}_n} \equiv f\left(\bar{X}_n\right).$$

Hence, setting

$$x \equiv \bar{X}_n \quad \textit{and} \quad x_0 \equiv \mu_{\bar{X}_n} \equiv \mathbf{E}\left[\bar{X}_n\right] = \mathbf{E}\left[X\right] = \frac{1}{\lambda},$$

the Taylor formula yields

$$\hat{\lambda}_n^M \approx f\left(\mu_{\bar{X}_n}\right) + f'\left(\mu_{\bar{X}_n}\right)\left(\bar{X}_n - \mu_{\bar{X}_n}\right) = \lambda - \lambda^2\left(\bar{X}_n - \frac{1}{\lambda}\right),$$

It follows

$$\mathbf{E}\left[\hat{\lambda}_n^M\right] \approx \lambda - \lambda^2\left(\mathbf{E}\left[\bar{X}_n\right] - \frac{1}{\lambda}\right) = \lambda$$

and

$$\mathbf{D}^2\left[\hat{\lambda}_n^M\right] \approx \mathbf{D}^2\left[f\left(\mu_{\bar{X}_n}\right) + f'\left(\mu_{\bar{X}_n}\right)\left(\bar{X}_n - \mu_{\bar{X}_n}\right)\right] = f'\left(\mu_{\bar{X}_n}\right)^2 \mathbf{D}^2\left[\bar{X}_n\right] = \lambda^4 \frac{1}{n\lambda^2} = \frac{\lambda^2}{n}.$$

As a consequence, we have that the estimator $\hat{\lambda}_n^M$ is approximatively unbiased. In addition, we can write

$$\mathbf{E}\left[\left(\hat{\lambda}_n^M - \lambda\right)^2\right] \approx \mathbf{E}\left[\left(\hat{\lambda}_n^M - \mathbf{E}\left[\hat{\lambda}_n^M\right]\right)^2\right] = \mathbf{D}^2\left[\hat{\lambda}_n^M\right] \approx \frac{\lambda^2}{n}.$$

This implies that

$$\lim_{n\to\infty} \mathbf{E}\left[\left(\hat{\lambda}_n^M - \lambda\right)^2\right] = \lim_{n\to\infty} \frac{\lambda^2}{n} = 0.$$

That is the estimator $\hat{\lambda}_n^M$ is consistent in mean square. A fortiori, the estimator $\hat{\lambda}_n^M$ is consistent in probability. Note that the latter claim can be proved directly by oserving that

$$\bar{X}_n \xrightarrow{\boldsymbol{P}} \mathbf{E}\left[X\right] \quad \textit{and} \quad \mathbf{P}\left(\bar{X}_n > 0\right) = 1.$$

In Then, follows

$$\hat{\lambda}_n^M = \frac{1}{\bar{X}_n} \xrightarrow{\boldsymbol{P}} \frac{1}{\mathbf{E}\left[X\right]} = \lambda.$$

Note also that writing the distribution of X in terms of the scale paramenter $\theta \equiv 1/\lambda$ and applying again the method of moments we clearly obtain

$$\mu_1'\left(\theta\right) = \mathbf{E}\left[X\right] = \theta.$$

It immediately follows

$$\hat{\theta}_n = \bar{X}_n.$$

In this case, we have

$$\mathbf{E}\left[\hat{\theta}_n\right] = \mathbf{E}\left[\bar{X}_n\right] = \mathbf{E}\left[X\right] = \theta.$$

Hence, the estimator $\hat{\lambda}$ is biased and $\hat{\theta}$ is unbiased.

Example 1130 *Assume that X is a normal random variable with mean parameter μ and variance parameter σ^2. To build estimators $\hat{\mu}_n$ and $\hat{\sigma}_n^2$ for μ and σ^2 by the method of moments we compute the first and second population raw moment. We have*

$$\mu_1'\left(\mu,\sigma^2\right)=\mathbf{E}\left[X\right]=\mu,\quad \textit{and}\quad \mu_2'\left(\mu,\sigma^2\right)=\mathbf{E}\left[X^2\right]=\mu^2+\sigma^2.$$

Thereafter, we replace $\hat{\mu}_n$ to μ and $\hat{\sigma}_n^2$ to σ^2 and equate the first and second population moment to the first and second sample moment. This yields

$$\hat{\mu}_n=\bar{X}_n,\quad \textit{and}\quad \hat{\mu}_n^2+\hat{\sigma}_n^2=\bar{X}_n^{(2)}.$$

lt clearly follows

$$\hat{\mu}_n=\bar{X}_n$$

and, on account of Equation (14.7),

$$\hat{\sigma}_n^2=\bar{X}_n^{(2)}-\hat{\mu}_n^2=\bar{X}_n^{(2)}-\bar{X}_n^2=\tilde{S}_{X,n}^2.$$

Note that the determined estimator $\hat{\mu}_n$ is unbiased and $\hat{\sigma}_n^2$ is biased.

Example 1131 *Assume that X is a gamma random variable with shape parameter α and scale parameter θ. To build estimators $\hat{\alpha}_n$ and $\hat{\theta}_n$ for α and θ by the method of moments we compute the first and second population raw moment. We have*

$$\mu_1'\left(\alpha,\theta\right)=\mathbf{E}\left[X\right]=\alpha\theta,\quad \textit{and}\quad \mu_2'\left(\alpha,\theta\right)=\mathbf{E}\left[X^2\right]=\alpha\left(1+\alpha\right)\theta^2.$$

Thereafter, we replace $\hat{\alpha}_n$ to α and $\hat{\theta}_n$ to θ and equate the first and second population moment to the first and second sample moment. This yields

$$\hat{\alpha}_n\hat{\theta}_n=\bar{X}_n \tag{14.8}$$

and

$$\hat{\alpha}_n\left(1+\hat{\alpha}_n\right)\hat{\theta}_n^2=\bar{X}_n^{(2)}. \tag{14.9}$$

Considering Equations (14.8) and (14.7), Equation (14.9) becomes

$$\hat{\alpha}_n\hat{\theta}_n^2=\bar{X}_n^{(2)}-\bar{X}_n^2=\tilde{S}_n^2. \tag{14.10}$$

Thus, dividing the two sides of (14.10) by the two sides of (14.8), we obtain

$$\hat{\theta}_n=\frac{\tilde{S}_n^2}{\bar{X}_n}.$$

Replacing the obtained $\hat{\theta}_n$ in Equation (14.8), we end up with

$$\hat{\alpha}_n=\frac{\bar{X}_n^2}{\tilde{S}_n^2}.$$

14.2 Maximum Likelihood

Given a probabilistic model depending on some parameters, that is a probability distribution of the outcomes of a random phenomenon depending on some parameters, a *likelihood function* is a function of the parameters of the model given the outcomes. In statistical inference, likelihood functions play a key role to introduce a method for estimating the parameters of the model. In this context, the word "likelihood" should be thought as a counterpart of "probability" depending on the difference of the roles of outcomes versus parameters. Probability is used before (*ex-ante*) outcomes are available, as a function of the possible outcomes, given fixed values of the parameters. Likelihood is used after (*ex-post*) outcomes are available, as a function of the possible parameters, given fixed values of the outcomes.

Let X be a discrete [resp. absolutely continuous] real random variable with density function $f_X : \mathbb{Z} \times \Theta \to \mathbb{R}$ [resp. $f_X : \mathbb{R} \times \Theta \to \mathbb{R}$] depending on a parameter $\theta \in \Theta \subseteq \mathbb{R}^M$, for some $M \in \mathbb{N}$. Let $n \in \mathbb{N}$ and let $X_1, \ldots, X_n$ be a simple random sample of size n drawn from X.

Definition 1132 *We call* likelihood function *of the sample $X_1, \ldots, X_n$ the joint density of $X_1, \ldots, X_n$ treated as a function of the entries $\theta_1, \ldots, \theta_M$ of the parameter θ while considering the realizations $x_1, \ldots, x_n$ of the sample $X_1, \ldots, X_n$ as they were fixed parameters. That is the function $\mathcal{L}_{X_1,\ldots,X_n} : \Theta \times \mathbb{Z}^n \to \mathbb{R}$ [resp. $\mathcal{L}_{X_1,\ldots,X_n} : \Theta \times \mathbb{R}^n \to \mathbb{R}$] given by*

$$\begin{aligned} \mathcal{L}_{X_1,\ldots,X_n}(\theta_1, \ldots, \theta_M; x_1, \ldots, x_n) &\overset{def}{=} \textstyle\prod_{k=1}^{n} f_X(x_k; \theta_1, \ldots, \theta_M), \\ &\forall (\theta_1, \ldots, \theta_M) \in \Theta, \quad (x_1, \ldots, x_n) \in \mathbb{Z}^n \text{ [resp. } (x_1, \ldots, x_n) \in \mathbb{R}^n\text{]}. \end{aligned} \tag{14.11}$$

Remark 1133 *Assume that X is a discrete real random variable. Then, for a fixed value $\theta_0 \equiv \left(\theta_1^{(0)}, \ldots, \theta_M^{(0)}\right)$ of the parameter θ the function $\mathcal{L}_{X_1,\ldots,X_n|\theta_0} : \mathbb{Z}^n \to \mathbb{R}$ given by*

$$\mathcal{L}_{X_1,\ldots,X_n|\theta_0}(x_1, \ldots, x_n) \overset{def}{=} \mathcal{L}_{X_1,\ldots,X_n}\left(\theta_1^{(0)}, \ldots, \theta_M^{(0)}; x_1, \ldots, x_n\right) \quad \forall (x_1, \ldots, x_n) \in \mathbb{Z}^n \tag{14.12}$$

yields the probability of any realization $x_1, \ldots, x_n$ of the sample $X_1, \ldots, X_n$ when the parameter θ takes the value θ_0. On the contrary, for a fixed a realization $x_1^{(0)}, \ldots, x_n^{(0)}$ of the sample $X_1, \ldots, X_n$ the function $\mathcal{L}_{X_1,\ldots,X_n|x_1^{(0)},\ldots,x_n^{(0)}} : \Theta \to \mathbb{R}$ given by

$$\mathcal{L}_{X_1,\ldots,X_n|x_1^{(0)},\ldots,x_n^{(0)}}(\theta) \overset{def}{=} \mathcal{L}_{X_1,\ldots,X_n}\left(\theta_1, \ldots, \theta_M; x_1^{(0)}, \ldots, x_n^{(0)}\right) \quad \forall (\theta_1, \ldots, \theta_M) \in \Theta, \tag{14.13}$$

returns how likely is the value θ of parameter when the realization $x_1^0, \ldots, x_n^0$ of the sample $X_1, \ldots, X_n$ occurs.

Example 1134 (Bernoulli Likelihood) *Let X be a Bernoulli random variable with success probability parameter $\theta \equiv p$ In symbols $X \sim Ber(p)$. Setting $\Theta \equiv (0, 1)$, we know that the density function $f_X : \mathbb{Z} \times \Theta \to \mathbb{R}_+$ of X is given by*

$$f_X(x; p) = (1 - p) \cdot 1_{\{0\}}(x) + p \cdot 1_{\{1\}}(x), \tag{14.14}$$

for every $x \in \mathbb{Z}$ and $p \in \Theta$. However, in order to determine the likelihood function of a simple random sample of size n drawn from X, it is more convenient to rewrite $f_X(x; p)$ in the form

$$f_X(x; p) = p^x (1 - p)^{1-x} 1_{\{0,1\}}(x), \tag{14.15}$$

for every $x \in \mathbb{Z}$ and $p \in \Theta$. Let $X_1, \ldots, X_n$ be a simple random sample of size n drawn from X. Then, the likelihood function of $X_1, \ldots, X_n$ is given by

$$\begin{aligned}\mathcal{L}_{X_1,\ldots,X_n}(p; x_1, \ldots, x_n) &= \prod_{k=1}^{n} p^{x_k}(1-p)^{1-x_k} 1_{\{0,1\}}(x_k) \\ &= p^{\sum_{k=1}^{n} x_k}(1-p)^{n-\sum_{k=1}^{n} x_k} 1_{\{0,1\}^n}(x_1, \ldots, x_n), \end{aligned} \tag{14.16}$$

for every $p \in \Theta$ and all realizations $x_1, \ldots, x_n$ of the sample $X_1, \ldots, X_n$. It is worth noting that fixing p close to 0 [resp. 1] the probability of occurrence of the sequence of outcomes $1, \ldots, 1$, given by the value

$$\mathcal{L}_{X_1,\ldots,X_n|p}(1, \ldots, 1) = p^n,$$

is very small [resp. high]. On the contrary, fixing the sequence of outcomes $1, \ldots, 1$ the same value

$$\mathcal{L}_{X_1,\ldots,X_n|1,\ldots,1}(p) = p^n$$

informs us that is is very unlikely [resp. likely] that the success probability parameter p might take a value close to 0 [resp. 1].

Example 1135 (Binomial Likelihood) *Let X be a binomial random variable with number of trials parameter m and success probability parameter $\theta \equiv p$. In symbols $X \sim Bin(m, p)$. We assume that the parameter m is known (fixed) while the parameter p is unknown (variable). Hence, setting $\Theta \equiv (0, 1)$, we know that the density function $f_X : \mathbb{N}_0 \times \Theta \to \mathbb{R}$ of X is given by*

$$f_X(x; p) = \sum_{j=0}^{m} \binom{m}{j} p^j (1-p)^{m-j} \cdot 1_{\{j\}}(x), \tag{14.17}$$

for every $x \in \mathbb{N}_0$ and $p \in \Theta$. However, in order to determine the likelihood function of a simple random sample of size n drawn from X, it is more convenient to rewrite $f_X(x; p)$ in the form

$$f_X(x; p) = \frac{m!}{(m-x)!x!} p^x (1-p)^{m-x} \cdot 1_{\{0,1,\ldots,m\}}(x), \tag{14.18}$$

for every $x \in \mathbb{N}_0$ every $p \in \Theta$. Let $X_1, \ldots, X_n$ be a simple random sample of size n drawn from X. Then, the likelihood function $\mathcal{L}_{X_1,\ldots,X_n} : \Theta \times \mathbb{Z}^n \to \mathbb{R}$ of the sample $X_1, \ldots, X_n$ is given by

$$\begin{aligned}\mathcal{L}_{X_1,\ldots,X_n}(p; x_1, \ldots, x_n) &= \prod_{k=1}^{n} \frac{m!}{(m-x_k)!x_k!} p^{x_k}(1-p)^{m-x_k} \cdot 1_{\{0,1,\ldots,m\}}(x_k) \\ &= \left(\prod_{k=1}^{n} \frac{m!}{(m-x_k)!x_k!}\right) p^{\sum_{k=1}^{n} x_k}(1-p)^{n\cdot m-\sum_{k=1}^{n} x_k} 1_{\{0,1,\ldots,m\}^n}(x_1, \ldots, x_n) \end{aligned} \tag{14.19}$$

for every $p \in \Theta$ and all realizations $x_1, \ldots, x_n$ of the sample $X_1, \ldots, X_n$. Again, we note that that fixing p close to 0 [resp. 1] the probability of occurrence of the sequence of outcomes $m, \ldots, m$, given by the value

$$\mathcal{L}_{X_1,\ldots,X_n|p}(m, \ldots, m) = p^{n\cdot m},$$

is very small [resp. high]. On the contrary, fixing the sequence $m, \ldots, m$ the same value

$$\mathcal{L}_{X_1,\ldots,X_n|m,\ldots,m}(p) = p^{n\cdot m}$$

informs us that is is very unlikely [resp. likely] that the success probability parameter p might take a value close to 0 [resp. 1].

Example 1136 (Poisson Likelihood) *Let X be a Poisson random variable with rate parameter $\theta \equiv \lambda$. In symbols $X \sim Poiss(\lambda)$. Setting $\Theta \equiv \mathbb{R}_+$, we know that the density function $f_X : \mathbb{N}_0 \times \Theta \to \mathbb{R}_+$ of X is given by*

$$f_X(x;\lambda) = \sum_{j=0}^{\infty} \frac{e^{-\lambda}\lambda^j}{j!} \cdot 1_{\{j\}}(x), \tag{14.20}$$

for every $x \in \mathbb{N}_0$ and $\lambda \in \Theta$. However, in order to determine the likelihood function of a simple random sample of size n drawn from X, it is more convenient to rewrite $f_X(x;\lambda)$ in the form

$$f_X(x;\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \cdot 1_{\mathbb{N}_0}(x), \tag{14.21}$$

for every $x \in \mathbb{N}_0$ and $\lambda \in \Theta$. Let $X_1, \ldots, X_n$ be a simple random sample of size n drawn from X. Then, the likelihood function of the sample $X_1, \ldots, X_n$ is given by

$$\mathcal{L}_{X_1,\ldots,X_n}(\lambda; x_1, \ldots, x_n) = \prod_{k=1}^{n} \frac{e^{-\lambda}\lambda^x}{x!} \cdot 1_{\mathbb{N}_0}(x) = \frac{e^{-n\lambda}}{\prod_{k=1}^{n} x_k!} \lambda^{\sum_{k=1}^{n} x_k} 1_{\mathbb{N}_0^n}(x_1, \ldots, x_n) \tag{14.22}$$

for every $\lambda \in \Theta$ and all realizations $x_1, \ldots, x_n$ of the sample $X_1, \ldots, X_n$. Recall that $\lambda \equiv \mathbf{E}[X]$. Fixing λ close to 0 [far from 0] the probability of obtaining the realization of the sequence $0, \ldots, 0$, given by the value

$$\mathcal{L}_{X_1,\ldots,X_n|\lambda}(0, \ldots, 0) = e^{-n\lambda_0},$$

is very high [resp. small]. On the contrary, fixing the sequence $0, \ldots, 0$ the same value

$$\mathcal{L}_{X_1,\ldots,X_n|0,\ldots,0}(\lambda) = e^{-n\lambda}$$

informs us that is is very likely [resp. unlikely] that the success probability parameter λ might take a value close to 0 [resp. far from 0].

Example 1137 (Gaussian Likelihood) *Let X be a Gaussian random variable with mean μ and variance σ^2, setting $\theta \equiv (\mu, \sigma^2)$ and $\Theta \equiv \mathbb{R} \times \mathbb{R}_+$, we know that the density function $f_X : \mathbb{R} \times \Theta \to \mathbb{R}_+$ is given by*

$$f_X\left(x; \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(x-\mu)^2}{2\sigma^2}\right),$$

for every $x \in \mathbb{R}$ and $(\mu, \sigma^2) \in \Theta$. Let $X_1, \ldots, X_n$ be a simple random sample of size n drawn from X. Then, the likelihood function of the sample $X_1, \ldots, X_n$ is given by

$$\begin{aligned} \mathcal{L}_{X_1,\ldots,X_n}\left(\mu, \sigma^2; x_1, \ldots, x_n\right) &= \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_k-\mu)^2}{2\sigma^2}\right) \\ &= \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{k=1}^{n}(x_k-\mu)^2\right), \end{aligned} \tag{14.23}$$

for every $(\mu, \sigma^2) \in \Theta$ and all realizations $x_1, \ldots, x_n$ of the sample $X_1, \ldots, X_n$.

Definition 1138 *We say that the functions* $\theta_{n,1}^{ML} : \mathbb{Z}^n \to \mathbb{R}, \ldots, \theta_{n,M}^{ML} : \mathbb{Z}^n \to \mathbb{R}$, *[resp.* $\theta_{n,1}^{ML} : \mathbb{R}^n \to \mathbb{R}, \ldots, \theta_{n,M}^{ML} : \mathbb{R}^n \to \mathbb{R}$*],* $\theta_{n,1}^{ML} \equiv \theta_{n,1}^{ML}(x_1, \ldots, x_n), \ldots, \theta_{n,M}^{ML} \equiv \theta_{n,M}^{ML}(x_1, \ldots, x_n)$ *are maximum likelihood estimates of* $\theta_1, \ldots, \theta_M$ *if they fulfill the inequality*

$$\mathcal{L}_{X_1,\ldots,X_n}\left(\theta_{n,1}^{ML}, \ldots, \theta_{n,M}^{ML}; x_1, \ldots, x_n\right) \geq \mathcal{L}_{X_1,\ldots,X_n}(\theta_1, \ldots, \theta_M; x_1, \ldots, x_n), \tag{14.24}$$
$$\forall (\theta_1, \ldots, \theta_M) \in \Theta, \quad (x_1, \ldots, x_n) \in \mathbb{N}_0^n \text{ [resp. } (x_1, \ldots, x_n) \in \mathbb{R}^n\text{]}.$$

That is to say $\theta_{n,1}^{ML}, \ldots, \theta_{n,M}^{ML}$ *solve the equation*

$$\left(\theta_{n,1}^{ML}, \ldots, \theta_{n,M}^{ML}\right) = \underset{(\theta_1,\ldots,\theta_M)\in\Theta}{\arg\max}\, \mathcal{L}_{X_1,\ldots,X_n}(\theta_1, \ldots, \theta_M; x_1, \ldots, x_n). \tag{14.25}$$

Definition 1139 *Given maximum likelihood estimates* $\theta_{n,1}^{ML}, \ldots, \theta_{n,M}^{ML}$ *of* $\theta_1, \ldots, \theta_M$ *the estimators*

$$\hat{\theta}_{n,1}^{ML} \equiv \theta_{n,1}^{ML}(X_1, \ldots, X_n), \ldots, \hat{\theta}_{n,M}^{ML} \equiv \theta_{n,M}^{ML}(X_1, \ldots, X_n)$$

are called maximum likelihood estimators (MLE) of $\theta_1, \ldots, \theta_M$.

It is often difficult to maximize the likelihood function $\mathcal{L}_{X_1,\ldots,X_n}(\theta_1, \ldots, \theta_M; x_1, \ldots, x_n)$ through a direct approach. However, this difficulty may be circumvented, provided the density function $f_X : \mathbb{Z} \times \Theta \to \mathbb{R}_+$ [resp. $f_X : \mathbb{R} \times \Theta \to \mathbb{R}_+$] are strictly strictly positive. In this case, we introduce the so called *log-lkelihood function* which may be for help.

Definition 1140 *We call* log-likelihood function *of the sample* $X_1, \ldots, X_n$ *the natural logarithm of the joint likelihood function, that is the function* $\log \mathcal{L}_{X_1,\ldots,X_n} : \Theta \times \mathbb{Z}^n \to \mathbb{R}$ *[resp.* $\log \mathcal{L}_{X_1,\ldots,X_n} : \Theta \times \mathbb{R}^n \to \mathbb{R}$*] given by*

$$\log \mathcal{L}_{X_1,\ldots,X_n}(\theta_1, \ldots, \theta_M; x_1, \ldots, x_n) \overset{def}{=} \ln\left(\mathcal{L}_{X_1,\ldots,X_n}(\theta_1, \ldots, \theta_M; x_1, \ldots, x_n)\right),$$
$$\forall (\theta_1, \ldots, \theta_M) \in \Theta, \quad (x_1, \ldots, x_n) \in \mathbb{Z}^n \text{ [resp. } (x_1, \ldots, x_n) \in \mathbb{R}^n\text{]}.$$

Remark 1141 *Clearly,*

$$\log \mathcal{L}_{X_1,\ldots,X_n}(\theta_1, \ldots, \theta_M; x_1, \ldots, x_n) = \textstyle\sum_{k=1}^n \ln\left(f_{X_k,}(x_k; \theta_1, \ldots, \theta_M)\right),$$

for every $(\theta_1, \ldots, \theta_M) \in \Theta$ *and every* $(x_1, \ldots, x_n) \in \mathbb{Z}^n$ *[resp.* $(x_1, \ldots, x_n) \in \mathbb{R}^n$*].*

The maximization of the log-likelihood function $\log \mathcal{L}_{X_1,\ldots,X_n}(\theta_1, \ldots, \theta_M; x_1, \ldots, x_n)$ it is usually easier. This leads to determine $\theta_{n,1}^{ML}, \ldots, \theta_{n,M}^{ML}$ which solve the equation

$$\left(\theta_{n,1}^{ML}, \ldots, \theta_{n,M}^{ML}\right) = \underset{(\theta_1,\ldots,\theta_M)\in\Theta}{\arg\max}\, \log \mathcal{L}_{X_1,\ldots,X_n}(\theta_1, \ldots, \theta_M; x_1, \ldots, x_n). \tag{14.26}$$

Nevertheless, since the logarithm is a strictly increasing function Equations (14.25) and (14.26) have the same solution.

Example 1142 (Bernoulli Likelihood) *With reference to Example (??), we know that the likelihood function of a simple random sample* $X_1, \ldots, X_n$ *of size* n *drawn from a Bernoulli random variable* X *with success probability* p *is given by*

$$\mathcal{L}_{X_1,\ldots,X_n}(p; x_1, \ldots, x_n) = p^{\sum_{k=1}^n x_k}(1-p)^{n-\sum_{k=1}^n x_k} 1_{\{0,1\}^n}(x_1, \ldots, x_n)$$

for every $p \in \Theta$ and every $(x_1, \ldots, x_n) \in \mathbb{N}_0^n$. Now if we choose $x_1 = \cdots = x_n = 1$. We obtain

$$\mathcal{L}_{X_1,\ldots,X_n}(p; 1, \ldots, 1) = p^n,$$

for every $p \in \Theta$, and the likelihood function turns out to be strictly increasing in p. Similarly, if we choose $x_1 = \cdots = x_n = 0$. We obtain

$$\mathcal{L}_{X_1,\ldots,X_n}(p; 1, \ldots, 1) = (1-p)^n,$$

for every $p \in \Theta$, and the likelihood function turns out to be strictly decreasing in p. Assume $n = 2k$ for some $k \in \mathbb{N}$ and choose $x_1 = \cdots = x_k = 0$, $x_{k+1} = \cdots = x_n = 1$. We obtain

$$\mathcal{L}_{X_1,\ldots,X_n}(\theta; x_1, \ldots, x_n) = p^k (1-p)^k,$$

for every $\theta \in \Theta$, and the likelihood function is not strictly monotone in p. To deal with this case, consider the derivative

$$\frac{d}{dp} p^k (1-p)^k = k p^{k-1} (1-p)^k - k p^k (1-p)^{k-1} = k p^{k-1} (1-p)^{k-1} (1-2p).$$

We clearly have

$$\frac{d}{dp} p^k (1-p)^k = 0, \quad p = \frac{1}{2}.$$

In addition,

$$\begin{aligned}\frac{d^2}{dp^2} p^k (1-p)^k &= k(k-1) p^{k-2} (1-p)^k - k^2 p^{k-1} (1-p)^{k-1} - k^2 p^{k-1} (1-p)^{k-1} + k(k-1) p^k (1-p)^{k-2} \\ &= k p^{k-2} (1-p)^{k-2} \left((k-1)(1-p)^2 - 2kp(1-p) + (k-1) p^2 \right).\end{aligned}$$

Thus,

$$\left.\frac{d^2}{dp^2} p^k (1-p)^k\right|_{p=1/2} = k \left(\frac{1}{2}\right)^{2(k-2)} \left((k-1) \left(\frac{1}{2}\right)^2 - 2k \left(\frac{1}{2}\right)^2 + (k-1) \left(\frac{1}{2}\right)^2 \right) = -2k \left(\frac{1}{2}\right)^{2(k-1)} < 0.$$

Therefore, the likelihood function attains its maximum at the point $p = 1/2$. Hence, when the size n of the sample is even and the realization of the sample is given by the sequence $x_1 = \cdots = x_k = 0$, $x_{k+1} = \cdots = x_n = 1$, or any other sequence $x_1, \ldots, x_n$ which has the same number of 0's and 1's, the most likely value of the parameter p is $1/2$.

Example 1143 (Gaussian Likelihood) *With reference to Example (**??**), we know that the likelihood function of a simple random sample $X_1, \ldots, X_n$ of size n drawn from a Gaussian random variable X with mean μ and variance σ^2 is given by*

$$\mathcal{L}_{X_1,\ldots,X_n}(\mu, \sigma^2; x_1, \ldots, x_n) = \frac{1}{\sqrt{2^n \pi^n} \sigma^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{k=1}^n (x_k - \mu)^2 \right),$$

for every $(\mu, \sigma^2) \in \Theta$ and every $(x_1, \ldots, x_n) \in \mathbb{R}^n$. Now, setting $\bar{x}_n \equiv \frac{1}{n} \sum_{k=1}^n x_k$, we can write

$$\begin{aligned}\sum_{k=1}^n (x_k - \mu)^2 &= \sum_{k=1}^n (x_k - \bar{x}_n + \bar{x}_n - \mu)^2 \\ &= \sum_{k=1}^n \left((x_k - \bar{x}_n)^2 + 2(x_k - \bar{x}_n)(\bar{x}_n - \mu) + (\bar{x}_n - \mu)^2 \right) \\ &= \sum_{k=1}^n (x_k - \bar{x}_n)^2 + \sum_{k=1}^n 2(x_k - \bar{x}_n)(\bar{x}_n - \mu) + \sum_{k=1}^n (\bar{x}_n - \mu)^2.\end{aligned}$$

On the other hand,

$$\begin{aligned}\sum_{k=1}^{n} 2\left(x_k - \bar{x}_n\right)\left(\bar{x}_n - \mu\right) &= 2\left(\bar{x}_n - \mu\right)\sum_{k=1}^{n} 2\left(x_k - \bar{x}_n\right) \\ &= 2\left(\bar{x}_n - \mu\right)\left(\sum_{k=1}^{n} x_k - \sum_{k=1}^{n} \bar{x}_n\right) \\ &= 2\left(\bar{x}_n - \mu\right)\left(n\bar{x}_n - n\bar{x}_n\right) \\ &= 0\end{aligned}$$

and

$$\sum_{k=1}^{n}\left(\bar{x}_n - \mu\right)^2 = n\left(\bar{x}_n - \mu\right)^2.$$

Therefore,

$$\sum_{k=1}^{n}\left(x_k - \mu\right)^2 = \sum_{k=1}^{n}\left(x_k - \bar{x}_n\right)^2 + n\left(\bar{x}_n - \mu\right)^2.$$

It follows,

$$\mathcal{L}_{X_1,\ldots,X_n}\left(\theta; x_1, \ldots, x_n\right) = \frac{1}{\sqrt{2^n\pi^n}\sigma^n}\exp\left(-\frac{1}{2\sigma^2}\left(\sum_{k=1}^{n}\left(x_k - \bar{x}_n\right)^2 + n\left(\bar{x}_n - \mu\right)^2\right)\right).$$

Now, thanks to the positivitity of the terms $\left(x_k - \bar{x}_n\right)^2$*, for* $k = 1, \ldots, n$*, and* $\left(\bar{x}_n - \mu\right)$*, it is immediate to realize that for every fixed* $\sigma \in \mathbb{R}_+$*, the likelihood function attains its maximum at the point* $\mu = \bar{x}_n$*.*

Chapter 15

Confidence Intervals

Given a point estimator $\hat{\theta} : \Omega \to \mathbb{R}$ for the (unknown) value of a parameter $\theta \in \Theta \subseteq \mathbb{R}$ of a probability distribution and a sample point $\omega_0 \in \Omega$, the values $\hat{\theta}(\omega_0)$ taken by the estimator on the sample point provides an estimate of the true value of θ but provides no information about the reliability of such an estimate. That is, the point estimate says nothing about how close $\hat{\theta}(\omega_0)$ might be to the true value of θ. A way to deal with this lack of information is to report an intere interval of plausible values in which we are "confident" the true value θ of the parameter might be, the so called *confidence interval*. The word "confidence" is to denominate the uncertainty in a context in which the word "probability" cannot be used: when dealing with a numerical interval, we cannot say that the true value of a parameter is inside such an interval with some probability. The true value of a parameter is either inside a numeric interval or it is not. On the other hand, since the true value is unknown, there is no way that we can be sure it is inside the numeric interval considered. Hence, the word "confidence" is used in place of "probability" to address the uncertainty of whether the true value of the paramenter is inside a numeric interval. The larger is the confidence associated to a numeric interval, the more likely the true value of the parameter is inside the interval.

15.1 Critical Values

Let $X : \Omega \to \mathbb{R}$ be a real random variable on a probability space $\Omega \equiv (\Omega, \mathcal{E}, \mathbf{P})$ with distribution function $F_X : \mathbb{R} \to \mathbb{R}$.

Definition 1144 *Given any $\alpha \in (0,1)$, we call the* lower *[resp.* upper*]* tail critical value *of level α of X and denote it by x_α^- [resp. x_α^+] the minimum [resp. maximum] α-quantile [resp. $(1-\alpha)$-quantile] of X (see Definition 533). That is*

$$x_\alpha^- \overset{def}{=} \check{x}_\alpha \quad [resp.\ x_\alpha^+ \overset{def}{=} \hat{x}_{1-\alpha}].$$

Remark 1145 *We have*

$$x_\alpha^- = \min_{x \in \mathbb{R}} \{\mathbf{P}(X \leq x) \geq \alpha \quad and \quad \mathbf{P}(X \geq x) \geq 1-\alpha\} \tag{15.1}$$

and

$$x_\alpha^+ = \max_{x \in \mathbb{R}} \{\mathbf{P}(X \leq x) \geq 1-\alpha \quad and \quad \mathbf{P}(X \geq x) \geq \alpha\}. \tag{15.2}$$

Equivalently,

$$x_{\alpha}^{-}=\min_{x\in\mathbb{R}}\left\{F_{X}\left(x\right)\geq\alpha\quad\text{and}\quad\lim_{u\rightarrow x^{-}}F_{X}\left(x\right)\leq\alpha\right\} \tag{15.3}$$

and

$$x_{\alpha}^{+}=\max_{x\in\mathbb{R}}\left\{F_{X}\left(x\right)\geq1-\alpha\quad\text{and}\quad\lim_{u\rightarrow x^{-}}F_{X}\left(x\right)\leq1-\alpha\right\}. \tag{15.4}$$

Remark 1146 *Assume F_X is strictly increasing and continuous. Then, x_{α}^{-} [resp. x_{α}^{+}] is the unique solution of the equations*

$$F_{X}\left(x\right)=\alpha\quad\text{[resp. }F_{X}\left(x\right)=1-\alpha\text{]}, \tag{15.5}$$

for every $\alpha\in(0,1)$.

Remark 1147 *Assume F_X is absolutely continuous with density function f_X. Then, x_{α}^{-} and x_{α}^{+} are the unique solutions of the equations*

$$\int_{-\infty}^{x}f_{X}\left(u\right)d\mu_{L}\left(u\right)=\alpha\quad\text{and}\quad\int_{x}^{+\infty}f_{X}\left(u\right)d\mu_{L}\left(u\right)=\alpha, \tag{15.6}$$

respectively.

Definition 1148 *Given any $\alpha\in(0,1)$, we call the* two-tailed critical values *of level α of X the upper and lower critical values of level $\alpha/2$. That is the real numbers $x_{\alpha/2}^{-}$ and $x_{\alpha/2}^{+}$ introduced in Definition 1144.*

Remark 1149 *Assume F_X is strictly increasing and continuous. Then, $x_{\alpha/2}^{-}$ and $x_{\alpha/2}^{+}$ are the unique solutions of the equations*

$$F_{X}\left(x\right)=\alpha/2\quad\text{and}\quad F_{X}\left(x\right)=1-\alpha/2, \tag{15.7}$$

respectively.

Remark 1150 *Assume F_X is absolutely continuous with density function f_X. Then, x_{α}^{-} and x_{α}^{+} are the unique solutions of the equations*

$$\int_{-\infty}^{x}f_{X}\left(u\right)d\mu_{L}\left(u\right)=\alpha/2\quad\text{and}\quad\int_{x}^{+\infty}f_{X}\left(u\right)d\mu_{L}\left(u\right)=\alpha/2, \tag{15.8}$$

respectively.

15.2 Confidence Bounds

Let $X:\Omega\rightarrow\mathbb{R}$ be a real random variable on a probability space $\Omega\equiv(\Omega,\mathcal{E},\mathbf{P})$ with distribution function $F_X:\mathbb{R}\times\Theta\rightarrow\mathbb{R}_{+}$ depending on a parameter $\theta\in\Theta\subseteq\mathbb{R}$.

Definition 1151 *Given any $\alpha\in(0,1)$, we say that a statistic $\underline{\theta}:\Omega\rightarrow\mathbb{R}$ is a* lower confidence bound *at the* confidence level $1-\alpha$ *or a* $100\left(1-\alpha\right)\%$ lower confidence bound *for θ if we have*

$$\mathbf{P}\left(\underline{\theta}\leq\theta\right)\geq1-\alpha, \tag{15.9}$$

for every $\theta\in\Theta$. If $\underline{\theta}:\Omega\rightarrow\mathbb{R}$ is a lower confidence bound *for θ, Then, we call a* realization of the lower confidence bound $\underline{\theta}$ *for θ any value $\underline{\theta}\left(\omega\right)\in\mathbb{R}$ taken by the statistic $\underline{\theta}$ on the occurence of a sample point $\omega\in\Omega$.*

Definition 1152 *Given any $\alpha \in (0,1)$, we say that a statistic $\bar{\theta} : \Omega \to \mathbb{R}$ is an* upper confidence bound *at the* confidence level $1-\alpha$ *or a* $100\,(1-\alpha)\,\%$ upper confidence bound *for θ if we have*

$$\mathbf{P}\left(\bar{\theta} \geq \theta\right) \geq 1-\alpha, \tag{15.10}$$

for every $\theta \in \Theta$. If $\bar{\theta} : \Omega \to \mathbb{R}$ is an upper confidence bound *for θ, Then, we call a* realization of the upper confidence bound *for θ any value $\bar{\theta}(\omega) \in \mathbb{R}$ taken by the statistic $\bar{\theta}$ on the occurence of a sample point $\omega \in \Omega$.*

Note that Equations (15.9) and (15.10) do not prevent that on the occurrence of some sample points $\bar{\omega}, \underline{\omega} \in \Omega$ we might have $\underline{\theta}(\bar{\omega}) > \theta_0$ or $\bar{\theta}(\underline{\omega}) < \theta_0$, where θ_0 is the true value of the parameter θ.

Example 1153 *Assume X is Gaussian distributed with unknown mean $\mu \in \mathbb{R}$ and known variance σ^2, let $X_1, \ldots, X_n$ be a simple random sample of size n drawn from X, and let $\bar{X}_n$ be the sample mean of size n drawn from X. Then, the statistics*

$$\bar{X}_n - \frac{\sigma}{\sqrt{n}} z_\alpha \quad \textit{and} \quad \bar{X}_n + \frac{\sigma}{\sqrt{n}} z_\alpha,$$

where $z_\alpha \equiv z_\alpha^+$ is the upper tail critical value of level α of the standard Gaussian random variable[1]*, are respectively a lower and an upper bound for the true value of the parameter μ at the confidence level $1-\alpha$. The real numbers*

$$\bar{x}_n - \frac{\sigma}{\sqrt{n}} z_\alpha \quad \textit{and} \quad \bar{x}_n + \frac{\sigma}{\sqrt{n}} z_\alpha,$$

where $\bar{x}_n \equiv \bar{X}_n(\omega)$ is the value taken by the sample mean estimator on the occurrence of a sample point $\omega \in \Omega$, are respectively a realization of the lower and upper bound for μ.

Discussion. Since X is Gaussian distributed, also the statistic $\bar{X}_n$ is Gaussian distributed with mean μ and variance σ^2/n. Therefore, the statistic $\left(\bar{X}_n - \mu\right)/\sigma/\sqrt{n}$ is standard Gaussian distributed. As a consequence, considering the upper critical value z_α^+ of the standard normal distribution, we have

$$\mathbf{P}\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq z_\alpha^+\right) \geq 1-\alpha. \tag{15.11}$$

On the other hand,

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq z_\alpha^+ \Leftrightarrow \bar{X}_n - z_\alpha^+ \sigma/\sqrt{n} \leq \mu. \tag{15.12}$$

Combining (15.11) and (15.12), it follows

$$\mathbf{P}\left(\bar{X}_n - \frac{\sigma}{\sqrt{n}} z_\alpha^+ \leq \mu\right) \geq 1-\alpha,$$

whatever is the true value of the parameter μ. In the end, setting $z_\alpha \equiv z_\alpha^+$, we have that $\bar{X}_n - \frac{\sigma}{\sqrt{n}} z_\alpha$ is a lower bound for μ at the confidence level $1-\alpha$. Similarly, considering the lower critical value z_α^- of the standard normal distribution, we have

$$\mathbf{P}\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \geq z_\alpha^-\right) \geq 1-\alpha. \tag{15.13}$$

[1] When dealing with confidence intervals it is customary to use the symbol z_α, rather than $z_{1-\alpha}$, to denote the $1-\alpha$ quantile of the standard Gaussian random variable. Correspondingly, the α quantile is denoted by $-z_\alpha$.

On the other hand,

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \geq z_\alpha^- \Leftrightarrow \bar{X}_n - z_\alpha^- \sigma/\sqrt{n} \geq \mu \tag{15.14}$$

Combining (15.13) and (15.14), it follows

$$\mathbf{P}\left(\bar{X}_n - \frac{\sigma}{\sqrt{n}} z_\alpha^- \geq \mu\right) \geq 1 - \alpha,$$

whatever is the true value of the parameter μ. Recalling that $z_\alpha^- = -z_\alpha^+$ and $z_\alpha^+ \equiv z_\alpha$ we have that $\bar{X}_n + \frac{\sigma}{\sqrt{n}} z_\alpha$ is an upper bound for μ at the confidence level $1 - \alpha$. □

A lower [resp. upper] confidence bound at the confidence level $1 - \alpha$ for the true value θ_0 of the parameter θ is just a random variable $\underline{\theta} : \Omega \to \mathbb{R}$ [resp. $\bar{\theta} : \Omega \to \mathbb{R}$] which fulfills the probabilistic Equation (15.9) [resp. (15.10)], whatever θ_0 might be, among all possible values in Θ. Before observing the value taken by the statistic $\underline{\theta}$ [resp. $\bar{\theta}$] on the occurrence of a sample point $\omega \in \Omega$, we can say that there is the probability $1 - \alpha$ that the observed value $\underline{\theta}(\omega)$ [resp. $\bar{\theta}(\omega)$] will be smaller [resp. larger] than θ_0. On trhe other hand, as already remarked, after observing $\underline{\theta}(\omega)$ [resp. $\bar{\theta}(\omega)$], we cannot talk any longer in terms of probability. The value $\underline{\theta}(\omega)$ [resp. $\bar{\theta}(\omega)$] will be either smaller [resp. larger] of θ_0 or not. Nevertheless, in principle, we ignore the numerical value of θ_0. Hence, we cannot check whether $\underline{\theta}(\omega) \leq \theta_0$ [resp. $\bar{\theta}(\omega) \geq \theta_0$] or not. As a solution to this puzzle we will say that we are "confident" at the $100(1 - \alpha)\%$ that the observed value $\underline{\theta}(\omega)$ [resp. $\bar{\theta}(\omega)$] of the statistic $\underline{\theta}$ [resp. $\bar{\theta}$] is actually smaller [resp. larger] than the θ_0. The term "confidence" should be interpreted in a "frequentist sense", meaning that if we observed a large number of values taken by the statistic $\underline{\theta}$ [resp. $\bar{\theta}$] on the realization of several sample points, Then, nearly the $100(1 - \alpha)\%$ of the observed values would be smaller [resp. larger] than θ_0.

Definition 1154 *Fixed any $\alpha \in (0, 1)$, we say that two statistics $\phi : \Omega \to \mathbb{R}$ and $\psi : \Omega \to \mathbb{R}$ constitute a* confidence interval *(ϕ, ψ) at the* confidence level *$1 - \alpha$ or a $100(1 - \alpha)\%$* confidence interval *for (the true value of) the parameter θ if we have*

$$\mathbf{P}(\phi \leq \theta \leq \psi) \geq 1 - \alpha, \tag{15.15}$$

for every $\theta \in \Theta$. If (ϕ, ψ) is a confidence interval for θ, Then, we call a realization of the confidence interval *for θ any interval $(\phi(\omega), \psi(\omega)) \subseteq \mathbb{R}$ where $\phi(\omega)$ and $\psi(\omega)$ are the values taken by ϕ and ψ, respectively, on the occurence of a sample point $\omega \in \Omega$.*

Note that, similarly to the case of the lower and upper bounds, Equation (15.15) does not prevent that, on the occurrence of some sample point $\bar{\omega}, \underline{\omega}, \omega^* \in \Omega$, we might have $\theta_0 < \phi(\bar{\omega}) \leq \psi(\bar{\omega})$ or $\phi(\underline{\omega}) \leq \psi(\underline{\omega}) < \theta_0$, where θ_0 is the true value of the parameter θ, or even $\psi(\omega^*) < \phi(\omega^*)$. In the first and second case, the respective realizations $(\phi(\bar{\omega}), \psi(\bar{\omega}))$ and $(\phi(\underline{\omega}), \psi(\underline{\omega}))$ of the confidence interval for θ will not contain θ_0; in the third case the realization $(\phi(\omega^*), \psi(\omega^*))$ of the confidence interval will degenerate to the empty set.

Remark 1155 *Fixed any $\alpha \in (0, 1)$, let $\underline{\theta} : \Omega \to \mathbb{R}$ and $\bar{\theta} : \Omega \to \mathbb{R}$ be a lower and an upper confidence bound, respectively, at the confidence level $1 - \alpha$ for the parameter θ. Then, the couple $\left(\underline{\theta}, \bar{\theta}\right)$ constitute a* confidence interval *at the* confidence level *$1 - \alpha$ for the parameter θ.*

Definition 1156 *Fixed any $\alpha \in (0, 1)$, let (ϕ, ψ) a confidence interval at the confidence level $1 - \alpha$ for the parameter θ. We call the* width *of the confidence interval the statistic $\psi - \phi$. We call the* precision *of the confidence interval the real number $1/\mathbf{E}[\psi - \phi]$.*

Example 1157 *With reference to Example 1153, fixed any $\alpha \in (0,1)$, the statistics*

$$\bar{X}_n - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \quad \textit{and} \quad \bar{X}_n + \frac{\sigma}{\sqrt{n}} z_{\alpha/2},$$

where $z_{\alpha/2} \equiv z^+_{\alpha/2}$ is the upper tail critical value of level $\alpha/2$ of the standard Gaussian random variable, constitute a confidence interval for the parameter μ at the confidence level $1-\alpha$.

Discussion. Recalling the discussion of Example 1153, we have

$$\mathbf{P}\left(z^-_{\alpha/2} \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq z^+_{\alpha/2} \right) \geq 1-\alpha. \tag{15.16}$$

where $z^-_{\alpha/2}$ [resp. $z^+_{\alpha/2}$] is the lower [resp. upper] tail critical values of level $\alpha/2$ of the standard Gaussian distribution. On the other hand,

$$\begin{aligned} z^-_{\alpha/2} \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq z^+_{\alpha/2} &\Leftrightarrow z^-_{\alpha/2}\sigma/\sqrt{n} \leq \bar{X}_n - \mu \leq z^+_{\alpha/2}\sigma/\sqrt{n} \\ &\Leftrightarrow -\bar{X}_n + z^-_{\alpha/2}\sigma/\sqrt{n} \leq -\mu \leq -\bar{X}_n + z^+_{\alpha/2}\sigma/\sqrt{n} \\ &\Leftrightarrow \bar{X}_n - z^+_{\alpha/2}\sigma/\sqrt{n} \leq \mu \leq \bar{X}_n - z^-_{\alpha/2}\sigma/\sqrt{n} \end{aligned} \tag{15.17}$$

Combining (15.16) and (15.17) It follows

$$\mathbf{P}\left(\bar{X}_n - \frac{\sigma}{\sqrt{n}} z^+_{\alpha/2} \leq \mu \leq \bar{X}_n - \frac{\sigma}{\sqrt{n}} z^-_{\alpha/2} \right) \geq 1-\alpha,$$

whatever is the true value of the parameter μ. Hence, the statistics $\bar{X}_n - \frac{\sigma}{\sqrt{n}} z^+_{\alpha/2}$ and $\bar{X}_n - \frac{\sigma}{\sqrt{n}} z^+_{\alpha/2}$ constitute a confidence interval for μ at the confidence level $1-\alpha$. Recalling that $z^-_{\alpha/2} = -z^+_{\alpha/2}$ and $z_{\alpha/2} \equiv z^+_{\alpha/2}$, the desired result immediately follows. □

Note that the choice of these statistics prevents the realizations of the confidence interval from being empty, yet cannot prevent such realizations from not containing the true value of the parameter.

15.3 Confidence Intervals for the Mean of a Population

Let X be a real random variable representing a population with unknown mean μ. Given any $n \in \mathbb{N}$, let $X_1, \ldots, X_n$ be a simple random sample of size n drawn from X, let $\bar{X}_n \equiv \frac{1}{n}\sum_{k=1}^n X_k$ [resp. $S_n^2 \equiv \frac{1}{n-1}\sum_{k=1}^n \left(X_k - \bar{X}_n\right)^2$ the sample mean [resp. unbiased sample variance] of X, and let $S_n \equiv \sqrt{S_n^2}$ be the unbiased sample standard deviation of X.

Proposition 1158 *Assume X is Gaussian distributed with known variance σ^2. Then, fixed any $\alpha \in (0,1)$, a $100\,(1-\alpha)\,\%$ confidence interval for (the true value of) the parameter μ (see Example 1157) is given by the statistics*

$$\bar{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \textit{and} \quad \bar{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \tag{15.18}$$

where $z_{\alpha/2} \equiv z^+_{\alpha/2}$ is the upper tail critical value of level $\alpha/2$ of the standard Gaussian random variable. The realizations of such a confidence interval are of the form

$$\left(\bar{x}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right), \tag{15.19}$$

where $\bar{x}_n$ is the value taken by the sample mean estimator $\bar{X}_n$ on any of the available realizations $x_1, \ldots, x_n$ of the sample $X_1, \ldots, X_n$.

Proof. Clearly the result can derived from the Discussion of Example 1157. As a direct proof, since under the assumption on X the statistic $\left(\bar{X}_n - \mu\right)/\sigma/\sqrt{n} \equiv Z$ has a standard Gaussian distribution, we have

$$\mathbf{P}\left(|Z| < z_{\alpha/2}\right) \geq 1 - \alpha \tag{15.20}$$

where $z_{\alpha/2}$ is the $\alpha/2$ upper tail critical value of $N(0,1)$. On the other hand,

$$\begin{aligned} |Z| < z_{\alpha/2} &\Leftrightarrow -z_{\alpha/2} < \frac{\mu - \bar{X}_n}{\sigma/\sqrt{n}} < z_{\alpha/2} \\ &\Leftrightarrow -z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu - \bar{X}_n < z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \\ &\Leftrightarrow \bar{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}. \end{aligned} \tag{15.21}$$

Combining (15.20) and (15.21) the desired result clearly follows. □

Proposition 1159 *With reference to Proposition 1158, fixed any $\alpha \in (0,1)$, the $100\,(1-\alpha)\,\%$ confidence interval for the true value of the parameter μ given by Equation (15.18) has a non-random width given by*

$$w = 2z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}.$$

Therefore, the precision of the interval is

$$\frac{1}{w} = \frac{\sqrt{n}}{2z_{\frac{\alpha}{2}}\sigma}.$$

As a consequence, we can achieve any desired precision provided we can choice the size n of the sample fulfilling the equation

$$n \geq \left(2z_{\frac{\alpha}{2}}\frac{\sigma}{w}\right)^2.$$

Proposition 1160 *Assume X is Gaussian distributed with unknown variance σ^2. Then, fixed any $\alpha \in (0,1)$, a $100\,(1-\alpha)\,\%$ confidence interval for the true value of the parameter μ is given by the statistics*

$$\bar{X}_n - t_{\frac{\alpha}{2},n-1}\frac{S_n}{\sqrt{n}} \quad \text{and} \quad \bar{X}_n + t_{\frac{\alpha}{2},n-1}\frac{S_n}{\sqrt{n}}, \tag{15.22}$$

where $t^+_{\frac{\alpha}{2},n-1}$ is the upper tail critical value of level $\alpha/2$ of the Student t-distribution with $n-1$ degrees of freedom. The realizations of such a confidence interval are of the form

$$\left(\bar{x}_n - t_{\frac{\alpha}{2},n-1}\frac{s_n}{\sqrt{n}}, \bar{x}_n + t_{\frac{\alpha}{2},n-1}\frac{s_n}{\sqrt{n}}\right), \tag{15.23}$$

where $\bar{x}_n$ [resp. s_n] is the value taken by the sample mean estimator $\bar{X}_n$ [resp. unbiased sample standard deviation estimator S_n] on any of the available realizations $x_1, \ldots, x_n$ of the sample $X_1, \ldots, X_n$.

Proof. Under the assumption on X, we know that the statistic $\left(\bar{X}_n - \mu\right)/S_n/\sqrt{n}$ has the Student distribution with $n-1$ degrees of freedom (see Theorem 1072). In symbols,

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \sim t_{n-1}.$$

Therefore, we have

$$\mathbf{P}\left(\left|\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}\right| < t_{n-1,\alpha/2}\right) \geq 1-\alpha,$$

where $t_{n-1,\alpha/2}$ is the $\alpha/2$ upper tail critical value of the Student distribution with $n-1$ degrees of freedom. To complete the proof is Then, sufficient to replicate the same argument used in the proof of Proposition 1158.

15.4 Approximate Confidence Intervals for the Mean of a Population

Let X be a real random variable representing a characteristic of the population with unknown mean μ. Given any $n \in \mathbb{N}$, let $X_1, \ldots, X_n$ be a simple random sample of size n drawn from X, let $\bar{X}_n \equiv \frac{1}{n}\sum_{k=1}^n X_k$ [resp. $S_n^2 \equiv \frac{1}{n-1}\sum_{k=1}^n \left(X_k - \bar{X}_n\right)^2$ the sample mean [resp. unbiased sample variance] of X and let $S_n \equiv \sqrt{S_n^2}$ be the "unbiased" sample standard deviation of X.

Proposition 1161 *Assume the distribution of X is unknown, but X has finite moment of order 4 and the size n of the sample $X_1, \ldots, X_n$ is large*[2]*. Then, fixed any $\alpha \in (0,1)$, an approximatively $100\,(1-\alpha)\,\%$ confidence interval for μ is given by the statistics*

$$\bar{X}_n - z_{\alpha/2}\frac{S_n}{\sqrt{n}} \quad \textit{and} \quad \bar{X}_n + z_{\alpha/2}\frac{S_n}{\sqrt{n}}, \tag{15.24}$$

where $z_{\alpha/2} \equiv z^+_{\alpha/2}$ is the upper tail critical value of level $\alpha/2$ of the standard Gaussian random variable. The realizations of such a confidence interval are of the form

$$\left(\bar{x}_n - z_{\alpha/2}\frac{s_n}{\sqrt{n}}, \bar{x}_n + z_{\alpha/2}\frac{s_n}{\sqrt{n}}\right), \tag{15.25}$$

where $\bar{x}_n$ [resp. s_n] is the value taken by the sample mean estimator $\bar{X}_n$ [resp. unbiases sample standard deviation estimator S_n] on any of the available realizations $x_1, \ldots, x_n$ of the sample $X_1, \ldots, X_n$.

Proof. Under the assumption on X and the size n of the sample, we know that the statistic $\left(\bar{X}_n - \mu\right)/S_n/\sqrt{n}$ is approximately Gaussian distributed (see Theorem 1074). Therefore, we have

$$\mathbf{P}\left(\left|\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}\right| < z_{\alpha/2}\right) \approx 1-\alpha,$$

where $z_{\alpha/2}$ is the $\alpha/2$ upper tail critical value of the standard Gaussian distribution. To complete the proof is Then, sufficient to replicate the same argument used in the proof of Proposition 1158. □

[2] As a rule of thumb, it is customary to consider large a size $n \geq 40$.

Proposition 1162 *Assume X is Bernoulli distributed with unknown success probability p and assume that the size n of the sample $X_1, \ldots, X_n$ is large*[3]*. Then, fixed any $\alpha \in (0,1)$, an approximatively $100\,(1-\alpha)\,\%$ confidence interval for p is given by the statistics*

$$\frac{\bar{X}_n + z^2_{\frac{\alpha}{2}}/2n - z_{\frac{\alpha}{2}}\sqrt{\frac{1}{n}\bar{X}_n\left(1-\bar{X}_n\right) + z^2_{\frac{\alpha}{2}}/4n^2}}{1 + z^2_{\frac{\alpha}{2}}/n} \quad \textit{and} \quad \frac{\bar{X}_n + z^2_{\frac{\alpha}{2}}/2n + z_{\frac{\alpha}{2}}\sqrt{\frac{1}{n}\bar{X}_n\left(1-\bar{X}_n\right) + z^2_{\frac{\alpha}{2}}/4n^2}}{1 + z^2_{\frac{\alpha}{2}}/n}, \tag{15.26}$$

where $z_{\alpha/2} \equiv z^+_{\alpha/2}$ is the upper tail critical value of level $\alpha/2$ of the standard Gaussian random variable. The realizations of such a confidence interval are of the form

$$\left(\frac{\bar{x}_n + z^2_{\frac{\alpha}{2}}/2n - z_{\frac{\alpha}{2}}\sqrt{\frac{1}{n}\bar{x}_n\left(1-\bar{x}_n\right) + z^2_{\frac{\alpha}{2}}/4n^2}}{1 + z^2_{\frac{\alpha}{2}}/n}, \frac{\bar{x}_n + z^2_{\frac{\alpha}{2}}/2n + z_{\frac{\alpha}{2}}\sqrt{\frac{1}{n}\bar{x}_n\left(1-\bar{x}_n\right) + z^2_{\frac{\alpha}{2}}/4n^2}}{1 + z^2_{\frac{\alpha}{2}}/n}\right), \tag{15.27}$$

where $\bar{x}_n$ is the value taken by the sample mean estimator $\bar{X}_n$ on any of the available realizations $x_1, \ldots, x_n$ of the sample $X_1, \ldots, X_n$.

Proof. By virtue of Theorem 1074, we know that the statistic $\left(\bar{X}_n - \mu\right)/S_n/\sqrt{n}$ is asymptotically standard Gaussian distributed. On the other hand,

$$\frac{\bar{X}_n - p}{S_n/\sqrt{n}} = \frac{\bar{X}_n - p}{\sqrt{p\,(1-p)\,/n}} \Big/ \frac{S_n}{\sqrt{p\,(1-p)}}$$

and

$$\frac{S_n}{\sqrt{p\,(1-p)}} \xrightarrow{\mathbf{P}} 1.$$

In fact,

$$\mathbf{E}\left[\frac{S_n^2}{p\,(1-p)}\right] = \frac{\mathbf{E}\left[S_n^2\right]}{p\,(1-p)} = \frac{\mathbf{D}^2\left[X\right]}{p\,(1-p)} = 1$$

and, since

$$\mathbf{D}^2\left[\frac{S_n^2}{p\,(1-p)}\right] = \frac{\mathbf{D}^2\left[S_n^2\right]}{p^2\,(1-p)^2} = \frac{1}{p^2\,(1-p)^2\,n}\left(\frac{1-3p\,(p-1)}{p\,(1-p)} - \frac{n-3}{n-1}\right),$$

by virtue of the Chebyshev inequality, we obtain

$$\frac{S_n^2}{p\,(1-p)} \xrightarrow{\mathbf{P}} 1.$$

Therefore, by virtue of the Slutsky Theorem 944, also

$$\frac{S_n}{\sigma} \xrightarrow{\mathbf{P}} 1.$$

As a consequence, it follows that

$$\frac{\bar{X}_n - p}{\sqrt{p\,(1-p)\,/n}} = \frac{\bar{X}_n - p}{S_n/\sqrt{n}} \frac{S_n}{\sqrt{p\,(1-p)}}$$

[3] As a rule of thumb, when X is Bernoulli distributed, it is customary to consider large a size n such that $\min\{np, nq\} \geq 10$.

is asymptotically Gaussian distributed. We can Then, write

$$\mathbf{P}\left(\left|\frac{\bar{X}_n - p}{\sqrt{p\left(1-p\right)/n}}\right| \leq z_{\frac{\alpha}{2}}\right) \gtrsim 1-\alpha.$$

In the end, since

$$\left(\bar{X}_n + z_{\frac{\alpha}{2}}^2/2n\right)^2 - \left(1 + z_{\frac{\alpha}{2}}^2/n\right)\bar{X}_n^2 = z_{\frac{\alpha}{2}}^2\left(\frac{1}{n}\bar{X}_n\left(1-\bar{X}_n\right) + z_{\frac{\alpha}{2}}^2/4n^2\right),$$

we have

$$\begin{aligned}
&\left|\frac{\bar{X}_n - p}{\sqrt{p\left(1-p\right)/n}}\right| \leq z_{\frac{\alpha}{2}} \\
&\Leftrightarrow \left|\bar{X}_n - p\right| \leq z_{\frac{\alpha}{2}}\sqrt{p\left(1-p\right)/n} \\
&\Leftrightarrow \left(\bar{X}_n - p\right)^2 \leq z_{\frac{\alpha}{2}}^2 p\left(1-p\right)/n \\
&\Leftrightarrow \left(1 + z_{\frac{\alpha}{2}}^2/n\right)p^2 - 2\left(\bar{X}_n + z_{\frac{\alpha}{2}}^2/2n\right)p + \bar{X}_n^2 \leq 0 \\
&\Leftrightarrow \frac{\bar{X}_n + z_{\frac{\alpha}{2}}^2/2n - z_{\frac{\alpha}{2}}\sqrt{\frac{1}{n}\bar{X}_n\left(1-\bar{X}_n\right) + z_{\frac{\alpha}{2}}^2/4n^2}}{1 + z_{\frac{\alpha}{2}}^2/n} \leq p \leq \frac{\bar{X}_n + z_{\frac{\alpha}{2}}^2/2n + z_{\frac{\alpha}{2}}\sqrt{\frac{1}{n}\bar{X}_n\left(1-\bar{X}_n\right) + z_{\frac{\alpha}{2}}^2/4n^2}}{1 + z_{\frac{\alpha}{2}}^2/n}.
\end{aligned}$$

The desired result clearly follows. □

15.5 Confidence Intervals for the Variance of a Population

Similarly to the case of the confidence intervals for the mean of a population, also for the variance it is possible to establish various results depending on what is known on the distribution of the population. We represent the population as a random variable X with unknown mean μ and variance σ^2. Furthermore, for any simple random sample $X_1, \ldots, X_n$ of size n drawn from X, we will consider the sample mean $\bar{X}_n \equiv \frac{1}{n}\sum_{k=1}^n X_k$ and unbiased variance $S_{X,n}^2 \equiv \frac{1}{n-1}\sum_{k=1}^n \left(X_k - \bar{X}_n\right)$.

15.5.1 Confidence Intervals for the Variance of a Gaussian Population

Assume X is Gaussian distributed. Then, independently of the size n of the sample, we know that the statistic $\left(n-1\right)S_{X,n}^2/\sigma^2$ has the chi-square distribution with $n-1$ degrees of freedom (see Theorem **??**). In symbols,

$$\frac{\left(n-1\right)S_{X,n}^2}{\sigma^2} \sim \chi_{n-1}^2.$$

Remark 1163 *Writing $\chi_{n-1,\alpha}^{2,-}$ [resp. $\chi_{n-1,\alpha}^{2,+}$] for the lower [resp. upper] critical value of level α of χ_{n-1}^2 we clearly have*

$$\chi_{n-1,\alpha}^{2,-} = \chi_{n-1,\alpha}^2 \quad \textit{[resp. } \chi_{n-1,\alpha}^{2,+} = \chi_{n-1,1-\alpha}^2\textit{]},$$

where $\chi^2_{n-1,\alpha}$ [resp. $\chi^2_{n-1,1-\alpha}$] is the α-quantile of χ^2_{n-1}. Hence, the lower [resp. upper] critical value of level α of χ^2_{n-1} satisfies the equation

$$\mathbf{P}\left(\chi^2_{n-1} \leq \chi^{2,-}_{n-1,\alpha}\right) = \int_0^{\chi^{2,-}_{n-1,\alpha}} f_{\chi^2_{n-1}}(x)\,dx = \alpha$$

$$\textit{[resp. } \mathbf{P}\left(\chi^2_{n-1} \geq \chi^{2,+}_{n-1,\alpha}\right) = \int_{\chi^{2,+}_{n-1,\alpha}}^{+\infty} f_{\chi^2_{n-1}}(x)\,dx = \alpha\textit{].}$$

where $f_{\chi^2_{n-1}} : \mathbb{R} \to \mathbb{R}$ is the density of χ^2_{n-1}, given by

$$f_{\chi^2_{n-1}}(x) \overset{def}{=} \frac{1}{2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)} x^{\frac{n-1}{2}-1} e^{-\frac{x}{2}} 1_{\mathbb{R}_+}(x), \quad \forall x \in \mathbb{R}$$

*(see Definition **??**).*

Remark 1164 *We have*

$$\chi^{2,-}_{n-1,\alpha} = \chi^{2,+}_{n-1,1-\alpha} \quad \textit{and} \quad \chi^{2,-}_{n-1,1-\alpha} = \chi^{2,+}_{n-1,\alpha},$$

for every $\alpha \in (0,1)$. However, since the chi-square distribution with n degrees of freedom is not symmetric,

$$\chi^{2,+}_{n-1,\alpha} \neq -\chi^{2,-}_{n-1,\alpha}.$$

Remark 1165 *We have*

$$\mathbf{P}\left(\chi^{2,-}_{n-1,\alpha/2} < \frac{(n-1)\,S^2_{X,n}}{\sigma^2} < \chi^{2,+}_{n-1,\alpha/2}\right) = 1-\alpha,$$

for every $\alpha \in (0,1)$.

Proposition 1166 *Assume X is Gaussian distributed with unknown variance σ^2. Then, fixed any $\alpha \in (0,1)$, a $100\,(1-\alpha)\,\%$ confidence interval for the parameter σ^2 is given by the statistics*

$$\frac{(n-1)\,S^2_{X,n}}{\chi^{2,+}_{n-1,\alpha/2}} \quad \textit{and} \quad \frac{(n-1)\,S^2_{X,n}}{\chi^{2,-}_{n-1,\alpha/2}}, \tag{15.28}$$

where $\chi^{2,-}_{n-1,\alpha/2}$ [resp. $\chi^{2,+}_{n-1,\alpha/2}$] is the lower [resp. upper] tail critical value of level $\alpha/2$ of χ^2_{n-1}. The realizations of such a confidence interval are of the form

$$\left(\frac{(n-1)\,s^2_{X,n}}{\chi^{2,+}_{n-1,\alpha/2}}, \frac{(n-1)\,s^2_{X,n}}{\chi^{2,-}_{n-1,\alpha/2}}\right) \tag{15.29}$$

where $s^2_{X,n}$ is the value taken by the unbiased sample variance estimator S^2_n on any of the available realizations $x_1, \ldots, x_n$ of the sample $X_1, \ldots, X_n$.

If the random variable X is not Gaussian the confidence intervals for the variance σ^2 of X determined above might be very far from valid, even for samples of large size n. In particular, if the distribution of X has thick tails. We will deal with the problem of determining confidence intervals for the variance σ^2 of a non-normal random variable in the next subsections.

15.5.2 Confidence Intervals for the Variance of Population from Large Samples

We know that

$$\mathbf{E}\left[S_{X,n}^2\right] = \sigma^2.$$

We also know that when X has finite moment of order four we have

$$\mathbf{D}^2\left[S_{X,n}^2\right] = \frac{1}{n}\left(\kappa - \frac{n-3}{n-1}\sigma^4\right) = \frac{\sigma^4}{n}\left(\hat{\kappa} - \frac{n-3}{n-1}\right),$$

where $\kappa \equiv \mu_4$ [resp. $\hat{\kappa} \equiv \hat{\mu}_4 \equiv \mu_4/\sigma^4$] is the kurtosis or central moment of order four, [resp. the standardazed kurtosis or standardized central moment of order four] of X (see Equation (13.29)). In addition, $S_{X,n}^2$ is asymptotically Gaussian distributed (see [**?**, Arnold, S.F. (1990), Mathematical Statistics, Prentice Hall, Upper Saddle ... Bain, L.J. (1978),]).

Proposition 1167 *Assume X has finite moment of order four. Then, fixed any $\alpha \in (0,1)$, an approximately $100\,(1-\alpha)\,\%$ confidence interval for the variance σ^2 of X, for a large sample size n, is given by the statistics*

$$\frac{S_{X,n}^2}{1 - z_{\alpha/2}\sqrt{\frac{1}{n}\left(\widetilde{Kurt}_{X,n} - 1\right)}} \quad \textit{and} \quad \frac{S_{X,n}^2}{1 + z_{\alpha/2}\sqrt{\frac{1}{n}\left(\widetilde{Kurt}_{X,n} - 1\right)}}, \tag{15.30}$$

*where $\widetilde{Kurt}_{X,n}$ is the biased standardized sample kurtosis of X (see Definition **??**) and $z_{\alpha/2}$ is the $\alpha/2$ upper tail critical value of $N(0,1)$. The realizations of such a confidence interval are of the form*

$$\left(\frac{s_{X,n}^2}{1 - z_{\alpha/2}\sqrt{\frac{1}{n}\left(\widetilde{kurt}_{X,n} - 1\right)}}, \frac{s_{X,n}^2}{1 + z_{\alpha/2}\sqrt{\frac{1}{n}\left(\widetilde{kurt}_{X,n} - 1\right)}}\right), \tag{15.31}$$

where $s_{X,n}^2$ [resp. $\widetilde{kurt}_{X,n}$] is the value taken by the sample variance estimator $S_{X,n}^2$ [resp. the biased standardized sample kurtosis estimator $\widetilde{Kurt}_{X,n}$] of X on any of the available realizations $x_1, \ldots, x_n$ of the sample $X_1, \ldots, X_n$.

Proof. By virtue of the large sample size assumption, the statistic $\left(S_{X,n}^2 - \sigma^2\right)/\sigma^2\sqrt{\frac{1}{n}\left(\widetilde{Kurt}_{X,n} - 1\right)}$ has approximately a standard Gaussian distribution. Therefore, we can write

$$\mathbf{P}\left(-z_{\alpha/2} \leq \frac{S_{X,n}^2 - \sigma^2}{\sigma^2\sqrt{\frac{1}{n}\left(\widetilde{Kurt}_{X,n} - 1\right)}} \leq z_{\alpha/2}\right) \simeq 1 - \alpha,$$

where $z_{\alpha/2}$ is the $\alpha/2$-critical value of $N(0,1)$. On the other hand,

$$
\begin{aligned}
&-z_{\alpha/2} \leq \frac{S_{X,n}^2 - \sigma^2}{\sigma^2 \sqrt{\frac{1}{n}\left(\widetilde{Kurt}_{X,n} - 1\right)}} < z_{\alpha/2} \\
&\Leftrightarrow -z_{\alpha/2} + \frac{1}{\sqrt{\frac{1}{n}\left(\widetilde{Kurt}_{X,n} - 1\right)}} \leq \frac{S_{X,n}^2}{\sigma^2 \sqrt{\frac{1}{n}\left(\widetilde{Kurt}_{X,n} - 1\right)}} \leq z_{\alpha/2} + \frac{1}{\sqrt{\frac{1}{n}\left(\widetilde{Kurt}_{X,n} - 1\right)}} \\
&\Leftrightarrow 1 - z_{\alpha/2}\sqrt{\frac{1}{n}\left(\widetilde{Kurt}_{X,n} - 1\right)} \leq \frac{S_{X,n}^2}{\sigma^2} \leq 1 + z_{\alpha/2}\sqrt{\frac{1}{n}\left(\widetilde{Kurt}_{X,n} - 1\right)} \\
&\Leftrightarrow \frac{S_{X,n}^2}{1 + z_{\alpha/2}\sqrt{\frac{1}{n}\left(\widetilde{Kurt}_{X,n} - 1\right)}} < \sigma^2 < \frac{S_{X,n}^2}{1 - z_{\alpha/2}\sqrt{\frac{1}{n}\left(\widetilde{Kurt}_{X,n} - 1\right)}}
\end{aligned}
$$

The desired result follows. □

15.6 Confidence Intervals for the Difference of the Means of Two Populations

Let $X : \Omega \to \mathbb{R}$ and $Y : \Omega \to \mathbb{R}$ two real random variables on a probability space $\Omega \equiv (\Omega, \mathcal{E}, \mathbf{P})$ representing two different populations with unknown mean μ_X and μ_Y, respectively, let $X_1, \ldots, X_m$ [resp. $Y_1, \ldots, Y_n$] be a simple random sample of size m [resp. n] drawn from X [resp. Y], for some $m \in \mathbb{N}$ [resp. $n \in \mathbb{N}$], and let $\bar{X}_m \equiv \frac{1}{m}\sum_{j=1}^{m} X_j$ [resp. $\bar{Y}_n \equiv \frac{1}{n}\sum_{k=1}^{n} Y_k$] be the sample mean of X [resp. Y].

Proposition 1168 *Assume the random variables X and Y are Gaussian distributed with known variances σ_X^2 and σ_Y^2, respectively. In addition, assume that the samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ are independent. Then, fixed any $\alpha \in (0,1)$, a $100(1-\alpha)\,\%$ confidence interval for the true value of the difference $\mu_X - \mu_Y$ is given by the statistics*

$$
\bar{X}_m - \bar{Y}_n - z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}} \quad \text{and} \quad \bar{X}_m - \bar{Y}_n + z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}, \tag{15.32}
$$

where $z_{\alpha/2} \equiv z_{\alpha/2}^{+}$ is the upper tail critical value of level $\alpha/2$ of the standard Gaussian random variable. The realizations of such a confidence interval are of the form

$$
\left(\bar{x}_m - \bar{y}_n - z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}, \bar{x}_m - \bar{y}_n + z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}\right), \tag{15.33}
$$

where $\bar{x}_m$ [resp. $\bar{y}_n$] is the value taken by the sample mean estimator $\bar{X}_m$ [resp. $\bar{Y}_n$] on any of the available realization $x_1, \ldots, x_m$ [resp. $y_1, \ldots, y_n$] of the sample $X_1, \ldots, X_m$ [resp. $Y_1, \ldots, Y_n$].

Proof. Under the assumptions on X and Y, the statistic $\bar{X}_m$ [resp. $\bar{Y}_n$] is Gaussian distributed with mean μ_X [resp. μ_Y] and variance σ_X^2/m [resp. σ_Y^2/n]. Moreover, the independence assumption on the samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ makes also $\bar{X}_m$ and $\bar{Y}_n$ independent. As

a consequence, the statistic $\bar{X}_m - \bar{Y}_n$ is Gaussian distributed with mean $\mu_X - \mu_Y$ and variance $\sigma_X^2/m + \sigma_Y^2/n$. Hence, the standardization of $\bar{X}_m - \bar{Y}_n$, that is the random variable

$$\frac{\bar{X}_m - \bar{Y}_n - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}} \equiv Z$$

has a standard Gaussian distribution. We can Then, write

$$\mathbf{P}\left(|Z| < z_{\alpha/2}\right) \geq 1 - \alpha,$$

where $z_{\alpha/2}$ is the $\alpha/2$ critical value of $N(0,1)$. On the other hand,

$$\begin{aligned}
|Z| < z_{\alpha/2} &\Leftrightarrow -z_{\alpha/2} < \frac{\mu_X - \mu_Y - \left(\bar{X}_m - \bar{Y}_n\right)}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}} < z_{\alpha/2} \\
&\Leftrightarrow -z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}} < \mu_X - \mu_Y - \left(\bar{X}_m - \bar{Y}_n\right) < z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}} \\
&\Leftrightarrow \bar{X}_m - \bar{Y}_n - z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}} < \mu_X - \mu_Y < \bar{X}_m - \bar{Y}_n + z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}.
\end{aligned}$$

This implies the desired result. □

Let $S_{X,m}^2 \equiv \frac{1}{m-1}\sum_{j=1}^m \left(X_j - \bar{X}_n\right)^2$ [resp. $S_{Y,n}^2 \equiv \frac{1}{n-1}\sum_{k=1}^n \left(Y_k - \bar{Y}_n\right)^2$] the unbiased sample variance estimator of size m [resp. n] drawn from X [resp. Y].

Definition 1169 *We call* pooled sample variance *the statistic S_p^2, given by*

$$S_p^2 \overset{def}{=} \frac{(m-1)\,S_{X,m}^2 + (n-1)\,S_{Y,n}^2}{m+n-2}. \tag{15.34}$$

Proposition 1170 *Assume the random variables X and Y are Gaussian distributed with the same unknown variance σ^2. In addition assume that the samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ are independent. Then, fixed any $\alpha \in (0,1)$, a $100\,(1-\alpha)\,\%$ confidence interval for $\mu_X - \mu_Y$ is given by the statistics*

$$\bar{X}_m - \bar{Y}_n - t_{\frac{\alpha}{2},m+n-2}\sqrt{S_p^2\left(\frac{1}{m} + \frac{1}{n}\right)} \quad \text{and} \quad \bar{X}_m - \bar{Y}_n + t_{\frac{\alpha}{2},m+n-2}\sqrt{S_p^2\left(\frac{1}{m} + \frac{1}{n}\right)}, \tag{15.35}$$

where $t_{\frac{\alpha}{2},m+n-2} \equiv t^+_{\frac{\alpha}{2},m+n-2}$ is the upper tail critical value of level $\alpha/2$ of the Student t-distribution with $m+n-2$ degree of freedom. The realizations of such a confidence interval are of the form

$$\left(\bar{x}_m - \bar{y}_n - t_{\frac{\alpha}{2},m+n-2}\sqrt{s_p^2\left(\frac{1}{m} + \frac{1}{n}\right)}, \bar{x}_m - \bar{y}_n + t_{\frac{\alpha}{2},m+n-2}\sqrt{s_p^2\left(\frac{1}{m} + \frac{1}{n}\right)}\right), \tag{15.36}$$

where $\bar{x}_m$ [resp. $\bar{y}_n$] is the value taken by $\bar{X}_m$ [resp. $\bar{Y}_n$] on any of the available realization $x_1, \ldots, x_m$ [resp. $y_1, \ldots, y_n$] of the sample $X_1, \ldots, X_m$ [resp. $Y_1, \ldots, Y_n$] and s_p^2 is the value taken by the pooled sample variance S_p^2 on the data set $x_1, \ldots, x_m$ [resp. $y_1, \ldots, y_n$] used to evaluate $\bar{x}_m$ [resp. $\bar{y}_n$].

Proof. Recall that if a random variable Z [resp. W] has the standard Gaussian distribution [resp. the χ^2 distribution with ν degrees of freedom, for some $\nu > 0$] and Z and W are independent, Then, the random variable $Z/\sqrt{W/\nu}$ has the Student distribution with ν degrees of freedom. Now, the independence and Gaussianity assumptions on the samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ imply that the statistic

$$\frac{\bar{X}_m - \bar{Y}_n - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}} \equiv Z$$

has a Gaussian distribution. Moreover, since

$$\mathbf{E}[Z] = \mathbf{E}\left[\frac{\bar{X}_m - \bar{Y}_n - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}}\right] = \frac{\mathbf{E}\left[\bar{X}_m\right] - \mathbf{E}\left[\bar{Y}_n\right] - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}} = 0$$

and

$$\mathbf{D}^2[Z] = \mathbf{D}^2\left[\frac{\bar{X}_m - \bar{Y}_n - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}}\right] = \frac{\mathbf{D}^2\left[\bar{X}_m\right] + \mathbf{D}^2\left[\bar{Y}_n\right]}{\sigma^2\left(\frac{1}{m} + \frac{1}{n}\right)} = \frac{\frac{\sigma^2}{m} + \frac{\sigma^2}{n}}{\sigma^2\left(\frac{1}{m} + \frac{1}{n}\right)} = 1.$$

Z is standard. Still, the independence and Gaussianity assumptions on the samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ imply that the statistics

$$\frac{(m-1)\,S^2_{X,m}}{\sigma^2} \quad \text{and} \quad \frac{(n-1)\,S^2_{Y,n}}{\sigma^2}$$

are chi-square distributed with $m-1$ and $n-1$ degree of freedom, respectively, and are also independent. Hence, the statistic

$$\frac{1}{\sigma^2}\left((m-1)\,S^2_{X,m} + (m-1)\,S^2_{Y,n}\right) \equiv W$$

is chi-square distributed with $m+m-2$ degrees of freedom. As a consequence, the statistic

$$\frac{Z}{\sqrt{W/(m+n-1)}} = \frac{\bar{X}_m - \bar{Y}_n - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}} \Big/ \sqrt{\frac{(m-1)\,S^2_m + (m-1)\,S^2_n}{\sigma^2\,(m+n-1)}} = \frac{\bar{X}_m - \bar{Y}_n - (\mu_X - \mu_Y)}{\sqrt{S^2_p\left(\frac{1}{m} + \frac{1}{n}\right)}}$$

is Student distributed with $m+m-2$ degrees of freedom. We can Then, write

$$\mathbf{P}\left(\left|\frac{\bar{X}_m - \bar{Y}_n - (\mu_X - \mu_Y)}{\sqrt{S^2_p\left(\frac{1}{m} + \frac{1}{n}\right)}}\right| \le t_{\frac{\alpha}{2},m+m-2}\right) \ge \alpha,$$

for any $\alpha \in (0,1)$. On the other hand

$$\left|\frac{\bar{X}_m - \bar{Y}_n - (\mu_X - \mu_Y)}{\sqrt{S^2_p\left(\frac{1}{m} + \frac{1}{n}\right)}}\right| \le t_{\frac{\alpha}{2},m+m-2}$$

$$\Leftrightarrow -t_{\alpha/2,m+n-2} \le \frac{\mu_X - \mu_Y - \left(\bar{X}_m - \bar{Y}_n\right)}{\sqrt{S^2_p\left(\frac{1}{m} + \frac{1}{n}\right)}} \le t_{\alpha/2,m+n-2}$$

$$\Leftrightarrow \bar{X}_m - \bar{Y}_n - t_{\alpha/2,m+n-2}\sqrt{S^2_p\left(\frac{1}{m} + \frac{1}{n}\right)} \le \mu_X - \mu_Y \le \bar{X}_m - \bar{Y}_n + t_{\alpha/2,m+n-2}\sqrt{S^2_p\left(\frac{1}{m} + \frac{1}{n}\right)}$$

and the desired (15.35) follows. □

Proposition 1171 *Assume the random variables X and Y are Gaussian distributed with unknown (different) variances σ_X^2 and σ_Y^2, respectively. In addition, assume that the samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ are independent. Then, fixed any $\alpha \in (0,1)$, an approximate $100\,(1-\alpha)\,\%$ confidence interval for $\mu_X - \mu_Y$. is given by the statistics*

$$\bar{X}_m - \bar{Y}_n - t_{\frac{\alpha}{2},\hat{\nu}} \sqrt{\frac{S_{X,m}^2}{m} + \frac{S_{Y,n}^2}{n}} \quad \text{and} \quad \bar{X}_m - \bar{Y}_n + t_{\frac{\alpha}{2},\hat{\nu}} \sqrt{\frac{S_{X,m}^2}{m} + \frac{S_{Y,n}^2}{n}}, \tag{15.37}$$

where $t_{\frac{\alpha}{2},\hat{\nu}} \equiv t^+_{\frac{\alpha}{2},\hat{\nu}}$ is the upper tail critical value of level $\alpha/2$ of the Student distribution with $\hat{\nu}$ degree of freedom, for

$$\hat{\nu} \equiv \left\lfloor \frac{\left(\frac{s_{X,m}^2}{m} + \frac{s_{Y,n}^2}{n} \right)^2}{\frac{s_{X,m}^4}{(m-1)m^2} + \frac{s_{Y,n}^4}{(n-1)n^2}} \right\rfloor, \tag{15.38}$$

where $\lfloor \cdot \rfloor : \mathbb{R} \to \mathbb{Z}$ is the floor function and $s_{X,m}^2$ [resp. $s_{Y,n}^2$] is the value taken by $S_{X,m}^2$ [resp. $S_{Y,n}^2$] on any of the available realization $x_1, \ldots, x_m$ [resp. $y_1, \ldots, y_n$] of the sample $X_1, \ldots, X_m$ [resp. $Y_1, \ldots, Y_n$]. The realizations of such a confidence interval are of the form

$$\left(\bar{x}_m - \bar{y}_n - t_{\frac{\alpha}{2},\hat{\nu}} \sqrt{\frac{s_{X,m}^2}{m} + \frac{s_{Y,n}^2}{n}}, \bar{x}_m - \bar{y}_n + t_{\frac{\alpha}{2},\hat{\nu}} \sqrt{\frac{s_{X,m}^2}{m} + \frac{s_{Y,n}^2}{n}} \right), \tag{15.39}$$

where $\bar{x}_m$ [resp. $\bar{y}_n$] is the value taken by $\bar{X}_m$ [resp. $\bar{Y}_n$] on the data set $x_1, \ldots, x_m$ [resp. $y_1, \ldots, y_n$] used to evaluate $s_{X,m}^2$ [resp. $s_{Y,n}^2$].

Proof. Recall that if a random variable Z [resp. W] has the standard Gaussian distribution [resp. the χ^2 distribution with ν degrees of freedom, for some $\nu > 0$] and Z and W are independent, Then, the random variable $Z/\sqrt{W/\nu}$ has the Student t distribution with ν degrees of freedom. Now, the independence and Gaussianity assumptions on the samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ imply that the statistic

$$\frac{\bar{X}_m - \bar{Y}_n - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}} \equiv Z$$

has a Gaussian distribution. Moreover, since

$$\mathbf{E}[Z] = \mathbf{E}\left[\frac{\bar{X}_m - \bar{Y}_n - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}} \right] = \frac{\mathbf{E}\left[\bar{X}_m\right] - \mathbf{E}\left[\bar{Y}_n\right] - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}} = 0$$

and

$$\mathbf{D}^2[Z] = \mathbf{D}^2\left[\frac{\bar{X}_m - \bar{Y}_n - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}} \right] = \frac{\mathbf{D}^2\left[\bar{X}_m\right] + \mathbf{D}^2\left[\bar{Y}_n\right]}{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}} = \frac{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}} = 1.$$

Z is standard. Hence, consider the statistic

$$W \equiv \frac{\frac{S_{X,m}^2}{m} + \frac{S_{Y,n}^2}{n}}{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}. \tag{15.40}$$

Since

$$\mathbf{E}\left[S^2_{X,m}\right] = \sigma^2_X \quad \text{and} \quad \mathbf{E}\left[S^2_{Y,n}\right] = \sigma^2_Y,$$

we have

$$\mathbf{E}\left[W\right] = \mathbf{E}\left[\frac{\frac{S^2_{X,m}}{m} + \frac{S^2_{Y,n}}{n}}{\frac{\sigma^2_X}{m} + \frac{\sigma^2_Y}{n}}\right] = \frac{\frac{\mathbf{E}\left[S^2_{X,m}\right]}{m} + \frac{\mathbf{E}\left[S^2_{Y,n}\right]}{n}}{\frac{\sigma^2_X}{m} + \frac{\sigma^2_Y}{n}} = 1. \tag{15.41}$$

Furthermore, since

$$\mathbf{D}^2\left[S^2_{X,m}\right] = \frac{\sigma^4_X}{m}\left(3 - \frac{m-3}{m-1}\right) = \frac{2\sigma^4_X}{m-1} \quad \text{and} \quad \mathbf{D}^2\left[S^2_{Y,n}\right] = \frac{\sigma^4_Y}{n}\left(3 - \frac{n-3}{n-1}\right) = \frac{2\sigma^4_Y}{n-1},$$

on account of the independence assumption on the samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$, we have

$$\mathbf{D}^2\left[W\right] = \mathbf{D}^2\left[\frac{\frac{S^2_{X,m}}{m} + \frac{S^2_{Y,n}}{n}}{\frac{\sigma^2_X}{m} + \frac{\sigma^2_Y}{n}}\right] = \frac{\frac{\mathbf{D}^2\left[S^2_{X,m}\right]}{m^2} + \frac{\mathbf{D}^2\left[S^2_{Y,n}\right]}{n^2}}{\left(\frac{\sigma^2_X}{m} + \frac{\sigma^2_Y}{n}\right)^2} = 2\frac{\frac{\sigma^4_X}{(m-1)m^2} + \frac{\sigma^4_Y}{(n-1)n^2}}{\left(\frac{\sigma^2_X}{m} + \frac{\sigma^2_Y}{n}\right)^2}. \tag{15.42}$$

In the end, note that the independence assumptions on the samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ implies that W is independent of Z. If W were χ^2-distributed with ν degrees of freedom, for some $\nu > 0$, we could invoke what recalled at the beginnig of the proof and conclude that the statistic

$$\frac{Z}{\sqrt{W/\nu}} = \frac{\bar{X}_m - \bar{Y}_n - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma^2_X}{m} + \frac{\sigma^2_Y}{n}}} \Big/ \sqrt{\frac{\frac{S^2_{X,m}}{m} + \frac{S^2_{Y,n}}{n}}{\frac{\sigma^2_X}{m} + \frac{\sigma^2_Y}{n}} \Big/ \nu} = \frac{\bar{X}_m - \bar{Y}_n - (\mu_X - \mu_Y)}{\sqrt{\left(\frac{S^2_{X,m}}{m} + \frac{S^2_{Y,n}}{n}\right)/\nu}}$$

has the Student t distribution with ν degrees of freedom. However, there is no $\nu > 0$, such that W is χ^2-distributed with ν degrees of freedom. In fact, in this case, we should have

$$\mathbf{E}\left[W\right] = \nu \quad \text{and} \quad \mathbf{D}^2\left[W\right] = 2\nu,$$

which would clearly contradict (15.41) and (15.42). However, it is possible to prove (see [**?**]) that, setting

$$\tilde{W} \equiv \nu W, \quad \text{where} \quad \nu \equiv \frac{\left(\frac{\sigma^2_X}{m} + \frac{\sigma^2_Y}{n}\right)^2}{\frac{\sigma^4_X}{(m-1)m^2} + \frac{\sigma^4_Y}{(n-1)n^2}},$$

and assuming m, n sufficiently large, for instance $\min\{m, n\} \geq 5$, the random variable $\tilde{W}$ is approximately χ^2-distributed with ν degrees of freedom. With reference to this, note that

$$\mathbf{E}\left[\tilde{W}\right] = \mathbf{E}\left[\nu W\right] = \nu \mathbf{E}\left[W\right] = \nu$$

and

$$\mathbf{D}^2\left[\tilde{W}\right] = \mathbf{D}^2\left[\nu W\right] = \nu^2 \mathbf{D}^2\left[W\right] = 2\frac{\left(\frac{\sigma^2_X}{m} + \frac{\sigma^2_Y}{n}\right)^4}{\left(\frac{\sigma^4_X}{(m-1)m^2} + \frac{\sigma^4_Y}{(n-1)n^2}\right)^2} \frac{\frac{\sigma^4_X}{(m-1)m^2} + \frac{\sigma^4_Y}{(n-1)n^2}}{\left(\frac{\sigma^2_X}{m} + \frac{\sigma^2_Y}{n}\right)^2} = 2\nu.$$

However, since

$$\frac{(m-1)\,S^2_{X,m}}{\sigma^2_X} \sim \chi^2_{m-1} = \Gamma\left(\frac{m-1}{2}, 2\right) \quad \text{and} \quad \frac{(n-1)\,S^2_{Y,n}}{\sigma^2_Y} \sim \chi^2_{n-1} = \Gamma\left(\frac{n-1}{2}, 2\right),$$

where we are using the scale parametrization of the gamma distribution, by virtue of the scaling property of the gamma distribution, we have

$$\delta\frac{S^2_{X,m}}{m} \sim \frac{2\delta\sigma^2_X}{m\,(m-1)}\Gamma\left(\frac{m-1}{2}, 2\right) = \Gamma\left(\frac{m-1}{2}, \frac{2\delta\sigma^2_X}{m\,(m-1)}\right)$$

and

$$\delta\frac{S^2_{Y,n}}{n} \sim \frac{\delta\sigma^2_Y}{n\,(n-1)}\Gamma\left(\frac{n-1}{2}, 2\right) = \Gamma\left(\frac{n-1}{2}, \frac{2\delta\sigma^2_Y}{n\,(n-1)}\right),$$

for any $\delta > 0$. Considering that $S^2_{X,m}$ and $S^2_{Y,n}$ are independent, it follows

$$\delta\left(\frac{S^2_{X,m}}{m} + \frac{S^2_{Y,n}}{n}\right) \sim \Gamma\left(\frac{m-1}{2}, \frac{2\delta\sigma^2_X}{m\,(m-1)}\right) + \Gamma\left(\frac{n-1}{2}, \frac{2\delta\sigma^2_Y}{n\,(n-1)}\right),$$

which cannot be chi-square distributed (see [**?**]). In particular,

$$\tilde{W} = \frac{\left(\frac{\sigma^2_X}{m} + \frac{\sigma^2_Y}{n}\right)^2}{\frac{\sigma^4_X}{(m-1)m^2} + \frac{\sigma^4_Y}{(n-1)n^2}} \frac{\frac{S^2_{X,m}}{m} + \frac{S^2_{Y,n}}{n}}{\frac{\sigma^2_X}{m} + \frac{\sigma^2_Y}{n}} = \frac{\frac{\sigma^2_X}{m} + \frac{\sigma^2_Y}{n}}{\frac{\sigma^4_X}{(m-1)m^2} + \frac{\sigma^4_Y}{(n-1)n^2}} \left(\frac{S^2_{X,m}}{m} + \frac{S^2_{Y,n}}{n}\right)$$

cannot be chi-square distributed. Nevertheless, as a consequence of what presented above, we obtain that the statistic

$$\frac{\bar{X}_m - \bar{Y}_n - (\mu_X - \mu_Y)}{\sqrt{\frac{S^2_{X,m}}{m} + \frac{S^2_{Y,n}}{n}}} = \frac{\bar{X}_m - \bar{Y}_n - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma^2_X}{m} + \frac{\sigma^2_Y}{n}}} \Big/ \sqrt{\nu\frac{\frac{S^2_{X,m}}{m} + \frac{S^2_{Y,n}}{n}}{\frac{\sigma^2_X}{m} + \frac{\sigma^2_Y}{n}}/\nu} = \frac{Z}{\sqrt{\tilde{W}/\nu}}$$

is approximately Student distributed with ν degrees of freedom. This implies that we can write

$$\mathbf{P}\left(\left|\frac{\bar{X}_m - \bar{Y}_n - (\mu_X - \mu_Y)}{\sqrt{\frac{S^2_m}{m} + \frac{S^2_n}{n}}}\right| < t_{\frac{\alpha}{2},\nu}\right) \approx \alpha, \tag{15.43}$$

for any $\alpha \in (0,1)$. On the other hand, since σ^2_X and σ^2_Y are unknown, also ν is unknown, but it is possible to show that replacing ν with the estimate $\hat{\nu}$ given by (15.38) Equation (15.43) still holds true. In the end, since

$$\begin{aligned}
&\left|\frac{\bar{X}_m - \bar{Y}_n - (\mu_X - \mu_Y)}{\sqrt{\frac{S^2_{X,m}}{m} + \frac{S^2_{Y,n}}{n}}}\right| < t_{\frac{\alpha}{2},\hat{\nu}} \\
&\Leftrightarrow -t_{\frac{\alpha}{2},\hat{\nu}} < \frac{\mu_X - \mu_Y - \left(\bar{X}_m - \bar{Y}_n\right)}{\sqrt{\frac{S^2_{X,m}}{m} + \frac{S^2_{Y,n}}{n}}} < t_{\frac{\alpha}{2},\hat{\nu}} \\
&\Leftrightarrow \bar{X}_m - \bar{Y}_n - t_{\frac{\alpha}{2},\hat{\nu}}\sqrt{\frac{S^2_{X,m}}{m} + \frac{S^2_{Y,n}}{n}} < \mu_X - \mu_Y < \bar{X}_m - \bar{Y}_n + t_{\frac{\alpha}{2},\hat{\nu}}\sqrt{\frac{S^2_{X,m}}{m} + \frac{S^2_{Y,n}}{n}}
\end{aligned}$$

it follows that an approximate $100\,(1-\alpha)\,\%$ confidence interval for $\mu_X - \mu_Y$ is given by Equation (15.37). □

Proposition 1172 *Assume the random variables X and Y have finite moment of order 2. In addition, assume that the samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ are independent. In the end, assume that the size of both the samples is large*[4]*. Then, given any $\alpha \in (0,1)$, an approximate $100\,(1-\alpha)\,\%$ confidence interval for $\mu_X - \mu_Y$ is given by the statistics*

$$\bar{X}_m - \bar{Y}_n - z_{\frac{\alpha}{2}} \sqrt{\frac{S^2_{X,m}}{m} + \frac{S^2_{Y,n}}{n}} \quad \text{and} \quad \bar{X}_m - \bar{Y}_n + z_{\frac{\alpha}{2}} \sqrt{\frac{S^2_{X,m}}{m} + \frac{S^2_{Y,n}}{n}}, \tag{15.44}$$

where $z_{\frac{\alpha}{2}} \equiv z^+_{\frac{\alpha}{2}}$ is the upper tail critical value of level $\alpha/2$ of the standard Gaussian distribution. The realizations of such a confidence interval are of the form

$$\left(\bar{x}_m - \bar{y}_n - z_{\frac{\alpha}{2}} \sqrt{\frac{s^2_{X,m}}{m} + \frac{s^2_{Y,n}}{n}}, \bar{x}_m - \bar{y}_n + z_{\frac{\alpha}{2},\hat{\nu}} \sqrt{\frac{s^2_{X,m}}{m} + \frac{s^2_{Y,n}}{n}} \right), \tag{15.45}$$

where $\bar{x}_m$ [resp. $\bar{y}_n$] is the value taken by $\bar{X}_m$ [resp. $\bar{Y}_n$] on any of the available realization $x_1, \ldots, x_m$ [resp. $y_1, \ldots, y_n$] of the sample $X_1, \ldots, X_m$ [resp. $Y_1, \ldots, Y_n$] and $s^2_{X,m}$ [resp. $s^2_{Y,n}$] is the value taken by $S^2_{X,m}$ [resp. $S^2_{Y,n}$] on the data set $x_1, \ldots, x_m$ [resp. $y_1, \ldots, y_n$] used to evaluate $\bar{x}_m$ [resp. $\bar{y}_n$].

Proof. . □

[4] A common rule of thumb is $\min\{m, n\} \geq 40$.

Chapter 16

Prediction Intervals

Let $X : \Omega \to \mathbb{R}$ be a real random variable on a probability space $\Omega \equiv (\Omega, \mathcal{E}, \mathbf{P})$ representing a population with distribution depending on unknown mean paramenter $\mu \in \mathbb{R}$. Let $X_1, \ldots, X_n$ be a simple random sample of size n drawn from X, for some $n \in \mathbb{N}$, let $\bar{X}_n \equiv \frac{1}{n}\sum_{k=1}^{n} X_k$ [resp. $S_n^2 \equiv \frac{1}{n-1}\sum_{k=1}^{n}\left(X_k - \bar{X}_n\right)^2$ the sample mean [resp. the unbiased sample variance] of size n drawn from X, and let $S_n \equiv \sqrt{S_n^2}$ be the unbiased sample standard deviation of size n drawn from X. Consider the possibility of drawing an additional sample X_{n+1} from X. We know that the sample mean is a point estimator for the mean parameter, that is $\bar{X}_n = \hat{\mu}$. We also know that

$$\mathbf{E}\left[X_{n+1} \mid X_1, \ldots, X_n\right] = \mathbf{E}\left[X_{n+1}\right] = \mu = \mathbf{E}\left[\bar{X}_n\right].$$

Therefore, we may think on $\bar{X}_n$ as predictor of the yet unobserved realization of X_{n+1}. On the other hand, the point prediction $\hat{\mu}$ conveys no information about reliability or precision. This leads to the idea of prediction interval, that, similarly to the case of confidence indervals, contitutes a statistical interval within which we are confident to observe the realization of X_{n+1}.

Proposition 1173 *Assume the random variable X is Gaussian distributed with known variance σ^2. Then, fixed any $\alpha \in (0,1)$, a $100\,(1-\alpha)\,\%$ prediction interval for the true value of the sample X_{n+1} is given by the statistics*

$$\bar{X}_n - z_{\alpha/2}\sigma\sqrt{1+\frac{1}{n}} \quad \textit{and} \quad \bar{X}_n + z_{\alpha/2}\sigma\sqrt{1+\frac{1}{n}}, \tag{16.1}$$

where $z_{\alpha/2} \equiv z^+_{\alpha/2}$ is the upper tail critical value of level $\alpha/2$ of the standard Gaussian random variable. The realizations of such a prediction interval are of the form

$$\left(\bar{x}_n - z_{\alpha/2}\sigma\sqrt{1+\frac{1}{n}}, \bar{x}_n + z_{\alpha/2}\sigma\sqrt{1+\frac{1}{n}}\right), \tag{16.2}$$

where $\bar{x}_n$ is the value taken by the sample mean estimator $\bar{X}_n$ on any of the available realizations $x_1, \ldots, x_n$ of the sample $X_1, \ldots, X_n$.

Proof. Under the assumption on X, the statistic $\left(X_{n+1} - \bar{X}_n\right)/\sigma\sqrt{1+\frac{1}{n}} \equiv Z$ has a standard Gaussian distribution. In fact, we clearly have

$$\mathbf{E}[Z] = \frac{1}{\sigma\sqrt{1+\frac{1}{n}}}\left(\mathbf{E}\left[X_{n+1}\right] - \mathbf{E}\left[\bar{X}_n\right]\right) = 0$$

and

$$\mathbf{D}^2[Z] = \frac{1}{\sigma^2 \frac{n+1}{n}} \left(\mathbf{D}^2[X_{n+1}] - \mathbf{D}^2\left[\bar{X}_n\right] \right) = \frac{1}{\sigma^2 \frac{n+1}{n}} \left(\sigma^2 + \frac{1}{n}\sigma^2 \right) = 1,$$

which show that Z is standardized. Moreover, the independence and Gaussianity of $X_1, \ldots, X_n$ imply that Z is Gaussian distributed. As a consequence, we can write

$$\mathbf{P}\left(|Z| < z_{\alpha/2} \right) \geq 1 - \alpha$$

where $z_{\alpha/2}$ is the upper tail critical value of level $\alpha/2$ of $N(0,1)$. On the other hand,

$$\begin{aligned} |Z| < z_{\alpha/2} &\Leftrightarrow -z_{\alpha/2} < \frac{X_{n+1} - \bar{X}_n}{\sigma\sqrt{1+\frac{1}{n}}} < z_{\alpha/2} \\ &\Leftrightarrow \bar{X}_n - z_{\alpha/2}\sigma\sqrt{1+\frac{1}{n}} < X_{n+1} < \bar{X}_n + z_{\alpha/2}\sqrt{1+\frac{1}{n}}. \end{aligned}$$

This implies the desired result. □

Proposition 1174 *Assume the random variable X is Gaussian distributed with unknown variance. Then, fixed any $\alpha \in (0,1)$, a $100(1-\alpha)\,\%$ prediction interval for the true value of the sample X_{n+1} is given by the statistics*

$$\bar{X}_n - t_{\alpha/2,n-1} S_{X,n}\sqrt{1+\frac{1}{n}} \quad \textit{and} \quad \bar{X}_n + t_{\alpha/2,n-1} S_{X,n}\sqrt{1+\frac{1}{n}}, \tag{16.3}$$

where $t_{\alpha/2,n-1} \equiv t^+_{\alpha/2,n-1}$ is the upper tail critical value of level $\alpha/2$ of the Student distribution with $n-1$ degrees of freedom. The realizations of such a prediction interval are of the form

$$\left(\bar{x}_n - t_{\alpha/2,n-1} s_n \sqrt{1+\frac{1}{n}}, \bar{x}_n + t_{\alpha/2,n-1} s_n \sqrt{1+\frac{1}{n}} \right) \tag{16.4}$$

where $\bar{x}_n$ [resp. s_n] is the value taken by the sample mean estimator $\bar{X}_n$ [resp. unbiased sample standard deviation estimator S_n] on any of the available realizations $x_1, \ldots, x_n$ of the sample $X_1, \ldots, X_n$.

Proof. The random variable $X_{n+1} - \bar{X}_n$ is Gaussian distributed with mean 0. Therefore, since

$$\mathbf{D}^2\left[X_{n+1} - \bar{X}_n\right] = \mathbf{D}^2[X_{n+1}] + \mathbf{D}^2\left[\bar{X}_n\right] = \sigma^2\left(1 + \frac{1}{n}\right),$$

the random variable $\left(X_{n+1} - \bar{X}_n\right) / \sigma\sqrt{1+\frac{1}{n}}$ is standard Gaussian distributed. On the other hand, we know that the random variable $(n-1)\,S^2_{X,n}/\sigma^2$ is chi-square distributed with $n-1$ degrees of freedom (see Theorem 1069). In addition, since $\bar{X}_n$ and $S^2_{X,n}$ are independent (see Theorem 1068) and X_{n+1} is independent of $X_1, \ldots, X_n$, which implies that also X_{n+1} and $S^2_{X,n}$ are independent, the random variables $\left(X_{n+1} - \bar{X}_n\right) / \sigma\sqrt{1+\frac{1}{n}}$ and $(n-1)\,S^2_{X,n}/\sigma^2$ are also independent. Applying Lemma 1070, we obtain that the statistic

$$\frac{\left(X_{n+1} - \bar{X}_n\right) / \sigma\sqrt{1+\frac{1}{n}}}{\sqrt{\frac{(n-1)S^2_{X,n}/\sigma^2}{n-1}}} = \frac{X_{n+1} - \bar{X}_n}{S_{X,n}\sqrt{1+\frac{1}{n}}}$$

has the Student distribution with $n-1$ degrees of freedom. As a consequence, we can write

$$\mathbf{P}\left(\left|\frac{X_{n+1}-\bar{X}_n}{S_{X,n}\sqrt{1+\frac{1}{n}}}\right| < t_{\alpha/2,n-1}\right) \geq 1-\alpha.$$

where $t_{\alpha/2,n-1}$ is the upper tail critical value of level $\alpha/2$ of the Student distribution with $n-1$ degrees of freedom. On the other hand,

$$\begin{aligned}\left|\frac{X_{n+1}-\bar{X}_n}{S_{X,n}\sqrt{1+\frac{1}{n}}}\right| < t_{\alpha/2,n-1} &\Leftrightarrow -t_{\alpha/2,n-1} < \frac{X_{n+1}-\bar{X}_n}{S_{X,n}\sqrt{1+\frac{1}{n}}} < t_{\alpha/2,n-1} \\ &\Leftrightarrow \bar{X}_n - t_{\alpha/2,n-1}S_{X,n}\sqrt{1+\frac{1}{n}} < X_{n+1} < \bar{X}_n - t_{\alpha/2,n-1}S_{X,n}\sqrt{1+\frac{1}{n}}.\end{aligned}$$

The desired result clearly follows. □

$$\frac{\bar{X}_n-\mu}{S_n/\sqrt{n}}$$

$$\frac{X_{n+1}-\bar{X}_n}{S_n/\sqrt{n}} = \frac{X_{n+1}+\mu-\bar{X}_n-\mu}{S_n/\sqrt{n}} = \frac{X_{n+1}-\mu}{S_n/\sqrt{n}} - \frac{\bar{X}_n-\mu}{S_n/\sqrt{n}}$$

Chapter 17

Hypothesis Testing

Often, the objective of a statistical investigation is not to estimate the true value of a parameter in the distribution of a population trait, but to decide which of two contradictory claims about the parameter is correct. This because we may have some reason to believe that the true value of a parameter is close to a precise threshold and we aim to know whether the estimate of the parameter justifies our belief or not. Methods to accomplish this task constitute the part of statistical inference known as *hypothesis testing*. Note that the point of view of hypothesis testing is nearly opposite to that of confidence interval. In case of confidence intervals, the goal is to determine the interval in which the true value of a parameter might be, given a statistical estimate of the parameter and an error tolerance. In case of hypothesis testing the goal is to determine whether an assumed value of the parameter might be true, given a statistical estimate of the parameter and an error tolerance.

The general idea of hypothesis testing involves:

1. making a *statistical hypothesis*, which is a testable claim either about the value of a parameter in a population distribution (e.g. the location, the scale, the shape, etc.) from which a sample is drawn or about the population distribution itself (e.g. normal, lognormal, logistic, etc.) or even about a property of the process which generates the population distribution (stationarity, autocorrelation, markovianity, etc.);

2. making a clear *alternative hypothesis*, which contradicts the statistical hypothesis;

3. setting an *error tolerance* of the statistical test;

4. introducing an appropriate *test statistic* to deal with the statistical hypothesis;

5. collecting a data sample;

6. evaluating the test statistic on the collected sample to decide whether to reject or not the statistical hypotesis in favor of the alternative, within the error tolerance.

Every hypothesis test requires the above steps. The test will be formulated so that the statistical hypothesis is initially preferred. Eventually, the statistical hypothesis will not be rejected in favor of the alternative hypothesis, unless the sample evidence strongly contradicts it, providing support to the alternative hypothesis.

A rather close analogy comes from the criminal trials held in the United States. The American judicial system assumes that "the defendant is innocent until proven guilty beyond a

reasonable doubt". This is the analogous of a statistical hypothesis testing, where the statistical hypothesis is the innocence presumption of the defendand and the alternative hypothesis is that the defendant is guilty. The DNA or fingerprint test or other tools of police investigation, such as wiretaps, search warrants, interviews of witnesses, might be considerd the analogous of the statistics used to test the statistical hypothesis. The data collection is the strict analogous of the collection of the criminal evidences, such as blood spots, hair samples, finger prints, tissue fibers, shoe prints. The statistic evaluation on the collected sample is the judicial trial itself. The innocence presumption is rejected or not by the jury in light of the results of the police investigation. The error tolerance may be represented by the formula "beyond a reasonable doubt". Note that the innocence presumption is the favored hypothesis. The burden of the proof is placed on the prosecution team, who must find "sufficient evidence" to make the assumption of innocence refutable in favor of guilt. Note also that, whether the jury rejects the innocence hypothesis or not, we do not prove that the defendant is guilty or innocent. We cannot be absolutely certain that the defendand is guilty or innocent. We just behave as if the defendant is guilty or innocent, but there may always be the possibility of a judicial error. The judicial errors are that an innocent is found guilty or a guilt is found innocent. Also in Statistics we have an analogous situation. We have to deal with two types of statistical errors for any collected data sample: a wrong rejection of the statistical hypothesis or a wrong non rejection.

Example 1175 *Students' rumors report that the average grade given by the nutty professors at an infamous exam of Probability and Statistics is* 20/30*. However, the pofessors assert that the true average grade is significantly higher than* 20/30*. How can the students test their claim against the professors' claim within a given error tolerance?*

Example 1176 *Consider a simple financial market made of a risk free asset, a stock, and a derivative on the stock. To build a simple mathematical model of this market it would be convenient that the* logarithmic returns *of the stock were normally distributed. How can a scholar test this claim, within a given error tolerance, against the alternative that the logarithmic returns of the stock are not normally distributed?*

Example 1177 *A researcher claims in a paper that the insurgence of autism in children is connected with the diffusion of the MMR vaccine. Eventually, he presents a linear regression, with a positive slope and a significantly high* R^2*, between the percentange of children of a certain country who were inoculated with the MMR vaccine in the last* 20 *years and the percentage of children diagnosed with autism in the same years. However, the referee of the paper objects that such a linear regression could be* spurious*, so that the resercher's claim could be false, because no evidence is provided that the variables under scrutiny are* stationary *or at least* cointegrated*. How can the researcher test for the stationarity or cointegration of the variables considered within a given error tolerance?*

Definition 1178 *We call the* null hypothesis*, commonly denoted by* H_0*, the claim that it is initially assumed to be true. The alternative hypothesis, which contradicts the null hypothesis is usually denoted by* H_1 *or* H_a*.*

The reason of the term *null hypothesis* originates from a rather legitimate conservative approach in scientific and technological research. A null hypothesis often expresses the current state of the knowledge on a experiment while the alternative hypothesis that a researcher would like to validate, sometimes referred to as the *research hypothesis*, expresses some form

of innovation in the knowledge. To abandon the current state of the knowledge in favor of an innovative state a strong support to the innovation has to be provided. The term *null* aims to suggest the idea of no change.

A statistical test is a method based on a suitable statistic and the sample data collected to decide whether the null hypothesis H_0 should be rejected in favor of the alterative hypothesis H_1 or not. A statistical test is focused on the null hypothesis. This will be rejected in favor of the alternative hypothesis, within a given error tolerance, only if the sample evidence suggest that the null hypothesis is false. Eventually, the only two possible conclusion from a statistical test are that we reject the null hypothesis in favor of the alternative or we fail to reject the null hypothesis in favor of the alternative. We stress once more that if we reject the null hypothesis, we do not prove that the null hypothesis is false and the alternative hypothesis is true. Similarly, if we do not reject the null hypothesis, we do not prove that the null hypothesis is true and the alternative is false. In either cases, the evidence provided by the sample leads us to behave as the null hypothesis were false or true. Indeed, we can identify two types of statistical errors.

Definition 1179 *We say that we commit an* error of the first type*, briefly a* type I error*, if we reject the null hypothesis when it is true. We say that we commit an* error of the second type*, briefly a* type II error*, if we do not reject the null hypothesis when it is false.*

Notation 1180 *The probability of committing a type I [resp. a type II] error in a statistical test is usually denoted by the greek letter α [resp. β]. That is,*

$$\alpha \equiv \mathbf{P}\left(\textit{reject } H_0 \mid H_0 \textit{ is true}\right) \quad \textit{[resp. } \beta \equiv \mathbf{P}\left(\textit{not reject } H_0 \mid H_0 \textit{ is false}\right)\textit{]}.$$

The notation makes clear that both the probability of committing a type I and a type II error in a statistical test are conditional probabilities.

Remark 1181 *To compute the probability of committing a type I [resp. type II] error in a statistical test we need to know the distribution of the test statistic conditional to the assumption that null hypothesis is true [resp. false].*

Remark 1182 *The probability of committing a type I error in a statistical test is the above mentioned error tolerance. Typically, α is chosen among* 0.10*,* 0.05*, and* 0.01*.*

Instead of demanding error-free statistical tests, which is impossible unless the data collected are from the entire population, the strategy should be to reduce the probability of making either types of error. A good statistical test is one in which the probability of making either types of error is small. However, we will see that trying to reduce the probability of a type I error in a statistical test, everything else being equal, necessarily implies an increasing of the probabilities of the corresponding type II error(s).

Remark 1183 *If we aim to reduce probability of committing a type I error α in a statistical test, we have to reduce the probability of the events leading to a rejection of H_0. As a consequence, everything else being equal, the probability of the events leading to a non rejection of H_0 will increase.*

Definition 1184 *Similarly to the definition of confidence interval, we call the* significance level *of a statistical test, the probability of committing a type I error α in statistical test when referred to an observed value of the test statistic on a data sample. More precisely, when the observed value of a test statistic will fall inside the event [resp. the complement of the event] whose occurrence allows to reject [resp. does not allow to reject] the null hypothesis H_0 with probability α, we will say that H_0 is rejected [resp. not rejected] at the significance level α.*

Example 1185 *It is known that, thanks to the bumper, an automobile model sustain no visible damage* 25% *of the time in a* 15 *Km/h crash tests. A modified bumper has been proposed to increase this percentage. How many cars equipped with the new bumper in a sample of size* $n = 20$ *should not substain visible damages in the crash test to certify an increased efficience of the new bumper? (see* **??***).*

Discussion. Let X be a random variable describing the result of the crash test for a single car. Considering as a success the occurrence of no visible damages, we can represent X as a standard Bernoulli with success probability $p_0 = 0.25$. In symbols, $X \sim Ber(p_0)$. If we replace the old bumper with the new bumper, a conservative approach to the innovation suggests that no improvement should be observed in terms of visible damages in the crash test. Therefore, the natural null hypothesis H_0 and the alternative hypothesis H_1 of our statistical test could be formulated as $H_0 : p \leq p_0$ and $H_1 : p > p_0$, respectively, where $p_0 \equiv 0.25$. On the other hand, it is rather evident that a reformulation of the null hypothesis as $H_0 : p = p_0$ allows to achieve the same goal while simplifying the computation. In fact, the rejection of $H_0 : p = p_0$ in favor of the alternative hypothesis $H_0 : p > p_0$ implies also the rejection of $H_0 : p \leq p_0$ and the failure to reject $H_0 : p = p_0$ implies itself the failure to reject $H_0 : p \leq p_0$.

Now, considering the results of the crash test for a number $n > 1$ of different cars, the structure of the crash test leads us to think that we are dealing with a simple random sample $X_1, \ldots, X_n$ of size n drawn from X. Hence, we are led to represent the number of cars substaining no visible damages in the crash test as the sample sum

$$Z_n = \sum_{k=1}^{n} X_k. \tag{17.1}$$

Choosing Z_n as our test statistic, under the null hypothesis $H_0 : p = p_0$, we know that Z_n is a binomial random variable with parameters n and p_0. In symbols, $Z_n \sim Bin\,(n, p_0)$. We recall that

$$Z_n = k, \quad \mathbf{P}\,(Z_n = k) = \binom{n}{k} p_0^k\,(1 - p_0)^{n-k}\,, \quad k = 0, 1, \ldots, n \tag{17.2}$$

and

$$\mathbf{E}\,[Z_n] = np_0. \tag{17.3}$$

First, we study the probability of committing a type I error

$$\alpha \equiv \mathbf{P}\,(\text{reject } H_0 \mid H_0 \text{ is true})\,. \tag{17.4}$$

Since our simple sample has size $n = 20$, under the null hypothesis $H_0 : p = p_0$, the expected number of cars substaining no visible damage in the crash test is given by

$$\mathbf{E}\,[Z_{20}] = np_0 = 20 \cdot 0.25 = 5. \tag{17.5}$$

However, thanks to the new bumper, we hope a significant inprovement in the results of the crash test. That is, we aim to reject H_0. Of course, we can never be sure that H_0 is false, but if H_0 were true, on the average 5 cars would substain no visible damage in the crash test. Hence, in a direction compatible with the alternative hypothesis, the larger the number of the cars with no visible damages in the crash test, the more unlikely H_0 is true. This is actually the best we can achieve. Therefore, we will reject H_0 when the number $m \leq n$ of cars with no visible damages in the sample collected is significantly larger than 5. Large enough to make

the null hypothesis unlikely. The question arises how to quantify the tems "large enough" and "unlikely"? On the other hand, assuming that H_0 is true, that is $Z_n \sim Bin(n, p_0)$, we can compute the minimum number m_α of cars with no visible damage to make the probability of committing a type I error equal to α, for any $\alpha \in (0,1)$. To this end, we have just to solve the equation

$$m_\alpha = \underset{0 \le m \le n}{\arg\min} \{\mathbf{P}(Z_n \ge m \mid H_0 \text{ is true})\} \le \alpha \tag{17.6}$$

As a consequence, choosing α significatively small to render the term "unlikely" meaningful in our context, we can determine the number m "large enough" to reject the null hypothesis as the solution of Equation (17.6). Now,

$$\mathbf{P}(Z_n \ge m \mid H_0 \text{ is true}) = \mathbf{P}(Z_n \ge m \mid Z_n \sim Bin(n, p_0)) = \sum_{k=m}^{n} \binom{n}{k} p_0^k (1-p_0)^{n-k}. \tag{17.7}$$

Therefore, choosing for instance $\alpha = 0.05$, we are lead to determine m_α as the smallest integer such that

$$\sum_{k=m}^{n} \binom{n}{k} p_0^k (1-p_0)^{n-k} \le 0.05. \tag{17.8}$$

We have

$$\sum_{k=m}^{n} \binom{n}{k} p_0^k (1-p_0)^{n-k} = \sum_{k=m}^{20} \binom{20}{k} 0.25^k \cdot 0.75^{20-k} \simeq \begin{cases} 0.102 & \text{for } m = 8 \\ 0.041 & \text{for } m = 9 \end{cases}. \tag{17.9}$$

It Then, clearly follows

$$m_\alpha = 9. \tag{17.10}$$

Rejecting H_0 if at least 9 cars in the sample of 20 substain no visible damage in the crash test with the new bumper yields a probability of committing a type I error lower than 0.05. Note that this number is significantly higher than the expected number of 5 cars computed if H_0 were true. Note also that ours is an "ex ante" computation. We have determined the minimum number m_α of cars with no visible damage to reject the null hypothesis with a probability of type I error lower than α, before considering the value taken by the test statistic on a collected data sample. After having observed the value of our test statistic on the collected data sample and having got the actual number m_0 of cars with no visible damage in the crash test, we cannot refer any longer to α as a probability. We have to refer to α as a *significance level*. We will say that H_0 is rejected [resp. not rejected] at the significance level α if m_0 is at least equal to [resp. smaller than] m_α.

To deal with the probability of committing a type II error requires a more detailed analysis. First, since we reject H_0 at the significance level $\alpha \equiv 0.05$ if at least 9 cars in the sample of 20 show no visible damage, we fail to reject H_0 at the significance lever α if not more than 8 cars in the sample of 20 substain no visible damage in the crash test with the new bumper. Second, conditioning to the hypothesis that H_0 is false, which means $p \neq p_0$ in the direction compatible with the alternative hypothesis, leads to a multiplicity of possible choices for p. Hence, there is a different value of β for each value of the parameter $p \neq p_0$ consistent with H_1. For instance, we could consider $p = p_1$ for $p_1 \equiv 0.3$, $p_1 \equiv 0.4$, $p_1 \equiv 0.5$, $p_1 \equiv 0.6$, $p_1 \equiv 0.7$, $p_1 \equiv 0.8$, etc. We Then, have a probability of committing a type II error which depends on the choice of p_1, (see

[**?**, Small-Sample Tests, p.453]). More specifically,

$$\beta(p_1) \equiv \mathbf{P}(\text{not reject } H_0 \mid p = p_1) = \sum_{k=0}^{8} \mathbf{P}(Z_n = k \mid p = p_1) = \sum_{k=0}^{8} \binom{20}{k} p_1^k (1-p_1)^{20-k}, \tag{17.11}$$

where

$$\sum_{k=0}^{8} \binom{20}{k} p_1^k (1-p_1)^{20-k} = \begin{cases} 0.887 & \text{for } p_1 = 0.3 \\ 0.596 & \text{for } p_1 = 0.4 \\ 0.252 & \text{for } p_1 = 0.5 \\ 0.057 & \text{for } p_1 = 0.6 \\ 0.005 & \text{for } p_1 = 0.7 \\ 0.000 & \text{for } p_1 = 0.8 \end{cases}. \tag{17.12}$$

Note that the probability $\beta(p_1)$ of accepting H_0 when H_0 is false, that is obtaining no more than 8 cars with no visible damages after the crash test, decreases quickly as p_1, expressing the proportion of the sample of n cars showing no damages in the crash test with the new bumper, grows away from p_0. Nevertheless, the probability of accepting H_0 when p_1 is slightly larger than p_0 is very high. Now, suppose that we aim to a more drastic reduction of the visible damage in the crash test, that is we try to render the significance level $\alpha \leq 0.01$. This leads to determine m such that

$$\sum_{k=m}^{20} \mathbf{P}(Z_n = k \mid H_0 \text{ is true}) = 0.01. \tag{17.13}$$

We have

$$\sum_{k=m}^{20} \binom{20}{k} 0.25^k \cdot 0.75^{20-k} \simeq \begin{cases} 0.014 & \text{for } m = 10 \\ 0.004 & \text{for } m = 11 \end{cases}. \tag{17.14}$$

Therefore, rejecting H_0 if at least 11 cars in the sample of 20 substain no visible damage in the crash test with the new bumper yields a probability of committing a type I error lower than 0.01. On the othe hand, in terms of the probability of a type II error we need to evaluate

$$\sum_{k=0}^{10} \binom{20}{k} p_1^k (1-p_1)^{20-k} = \begin{cases} 0.983 & \text{for } p_1 = 0.3 \\ 0.872 & \text{for } p_1 = 0.4 \\ 0.588 & \text{for } p_1 = 0.5 \\ 0.245 & \text{for } p_1 = 0.6 \\ 0.048 & \text{for } p_1 = 0.7 \\ 0.003 & \text{for } p_1 = 0.8 \end{cases}, \tag{17.15}$$

which significantly increase with respect to the corresponding probabilities computed with reference to the significance level $\alpha = 0.05$. □

Example 1186 *The drying time of a paint under specific test conditions is known to be normally distributed with mean value $\mu_0 \equiv 75$ min and standard deviation $\sigma_0 \equiv 9$ min. A chemist has proposed a new additive to decrease the average drying time, while the drying time remain normally distributed with the same standard deviation. Because of the investment which is necessary to introduce the new additive in the production process, a strong evidence should be provided that the new additive significantly decrease the average drying time. It is collected a sample of drying time of size $n = 25$ from different test specimens. Determine the average drying time of the specimens that leads to think on a significative decreasing of the former average drying time μ_0 (see [**?**, Example 9.2, p.430]).*

Discussion. Let X a random variable describing the drying time of a specimen of the paint under the test conditions. We can represent X as a normally distributed real random variable with mean μ and variance σ^2, that is $X \sim N(\mu, \sigma^2)$, where $\mu = \mu_0$ and $\sigma = \sigma_0$. A conservative approach to the innovation represented by the use of the new additive suggests that no inprovement should be observed in terms of the drying time. Therefore, the null hypothesis H_0 and the alternative hypothesis H_1 of our statistical test could be formulated as $H_0 : \mu \geq \mu_0$ and $H_1 : \mu < \mu_0$, respectively. However, the same argument exploited in Example 1185 allows to state the null hypothesis as $H_0 : \mu = \mu_0$. Considering a simple random sample $X_1, \ldots, X_n$ of size n drawn from X, we are Then, led to introduce the test statistic

$$Z_0 = \frac{\bar{X}_n - \mu_0}{\sigma_0/\sqrt{n}}. \tag{17.16}$$

This is convenient because Z_0 is normally distributed with variance 1. Moreover assuming H_0 true, Z_0 has null mean, $Z_0 \sim N(0, 1)$. Note that Z_0 expresses the distance between $\bar{X}_n$ and its expected value μ_0 when H_0 is true, in terms of standard deviation units. Hence, we can easily compute

$$\alpha \equiv \mathbf{P}\left(\text{reject } H_0 \mid H_0 \text{ is true}\right). \tag{17.17}$$

In fact, rejecting H_0, which is possible when $\bar{X}_n$ is significantly far from μ_0, is equivalent to reject H_0 when the realization $Z_0(\omega)$ of the statistic Z_0 on the specimen of the paint is significantly far from 0 in the direction compatible with the alternative hypothesis, that is a quite negative value. Setting a probability of type I error $\alpha \equiv 0.01$ and we know that for $z^-_{0.01} \equiv -2.33$ we have

$$\mathbf{P}\left(Z_0 < z\right) \leq 0.01, \tag{17.18}$$

for every $z < z^-_{0.01}$. Therefore, to reject the null hypothesis at a significance level $\alpha \equiv 0.01$ we need to obtain an observed value of our statistic Z_0 lower than $z^-_{0.01}$. This implies

$$Z_0(\omega) = \frac{\bar{X}_n(\omega) - \mu_0}{\sigma_0/\sqrt{n}} < z^-_{0.01} \tag{17.19}$$

That is

$$\bar{X}_n(\omega) < 75 - 2.33 \cdot \frac{9}{5} = 70.806. \tag{17.20}$$

Hence, we can reject the null hypothesis at a significance level $\alpha \equiv 0.01$ if the average drying time $\bar{X}_n(\omega)$ of the specimens turns out to be not larger 70.806 min. Equivalently, if we observe $Z_0(\omega) < z^-_{0.01}$. To study the probability of committing a type II error, we consider that we fail to reject H_0 at significance level $\alpha \equiv 0.01$ when $Z_0(\omega) \geq z^-_{0.01}$, so that the value $z^-_{0.01}$ is just the boundary betwen the rejection region and the not rejection region. Furthermore, H_0 is false when $\mu \neq \mu_0$ in the direction compatible with the alternative hypothesis. For instance, it might be possible $\mu = \mu_1$ for $\mu_1 \equiv 74$, $\mu_1 \equiv 73$, $\mu_1 \equiv 72$, $\mu_1 \equiv 71$, $\mu_1 \equiv 70$, $\mu_1 \equiv 69$, etc. Now, assuming $\mu = \mu_1$ true for a mispecified μ_1 implies that the statistic

$$Z_1 \equiv \frac{\bar{X}_n - \mu_1}{\sigma_0/\sqrt{n}} \tag{17.21}$$

is normally distributed with mean 0 and variance 1. On the other hand, since we can write

$$\frac{\bar{X}_n - \mu_0}{\sigma_0/\sqrt{n}} = \frac{\bar{X}_n - \mu_1 + \mu_1 - \mu_0}{\sigma_0/\sqrt{n}} = \frac{\bar{X}_n - \mu_1}{\sigma_0/\sqrt{n}} + \frac{\mu_1 - \mu_0}{\sigma_0/\sqrt{n}} \tag{17.22}$$

we have

$$Z_0 = Z_1 + \frac{\mu_1 - \mu_0}{\sigma_0/\sqrt{n}}. \tag{17.23}$$

It follows,

$$\begin{aligned}
\beta(\mu_1) &\equiv \mathbf{P}(\text{not reject } H_0 \mid \mu = \mu_1) = \mathbf{P}\left(Z_0 > z^-_{0.01} \mid \mu = \mu_1\right) \\
&= \mathbf{P}\left(Z_1 + \frac{\mu_1 - \mu_0}{\sigma_0/\sqrt{n}} > z^-_{0.01} \mid \mu = \mu_1\right) \\
&= \mathbf{P}\left(Z_1 > z^-_{0.01} - \frac{\mu_1 - \mu_0}{\sigma_0/\sqrt{n}} \mid \mu = \mu_1\right) \\
&= 1 - \Phi\left(z^-_{0.01} - \frac{\mu_1 - \mu_0}{\sigma_0/\sqrt{n}}\right).
\end{aligned} \tag{17.24}$$

We have

$$z^-_{0.01} - \frac{\mu_1 - \mu_0}{\sigma_0/\sqrt{n}} = \begin{cases} -1.774 & \text{for } \mu_1 = 74 \\ -1.219 & \text{for } \mu_1 = 73 \\ -0.663 & \text{for } \mu_1 = 72 \\ -0.108 & \text{for } \mu_1 = 71 \\ 0.448 & \text{for } \mu_1 = 70 \\ 1.003 & \text{for } \mu_1 = 69 \end{cases}. \tag{17.25}$$

Therefore,

$$\beta(\mu_1) = \begin{cases} 0.962 & \text{for } \mu_1 = 74 \\ 0.889 & \text{for } \mu_1 = 73 \\ 0.746 & \text{for } \mu_1 = 72 \\ 0.543 & \text{for } \mu_1 = 71 \\ 0.327 & \text{for } \mu_1 = 70 \\ 0.158 & \text{for } \mu_1 = 69 \end{cases}. \tag{17.26}$$

The probability of committing a type II error ir rather high if compared to the probability of committing a type I error. Increasing α from 0.01 to 0.05, would increase the probability of committing a type I error in a tolerable way while the probability of committing a type II error would be reduced. □

In what follows, the null hypothesis H_0 will be generally stated as an equality concerning a real parameter $\theta \in \Theta$ of a population distribution. In symbols we will write

$$H_0 : \theta = \theta_0,$$

where $\theta_0 \in \Theta$ is a given value, known as the *null value* of the parameter θ. The clear meaning of the above notation is that the null hypothesis consists in the guess of a specific value for the parameter θ. Hence, the alternative hypothesis H_1 will be stated as an inequality concerning θ and will be written in one of the possible forms

$$H_1 : \theta < \theta_0, \quad H_1 : \theta > \theta_0, \quad H_1 : \theta \neq \theta_0,$$

each of which contradicts H_0 is a different sense. Note that in many cases the combination of a null hypothesis stated as an equality and a suitable alternative hypothesis allows to deal with a wider generality while maintaining the technical simplicity. For instance, as already observed in Example 1185, testing the null hypothesis $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$ is equivalent to testing the null hypothesis $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$.

As we evaluate the statistic considered on the sample data collected, it is either unlikely or likely that the statistic would take the value it does if the null hypothesis were true. If it is unlikely, Then, we reject the null hypothesis in favor of the alternative hypothesis. If it is likely, Then, we do not reject the null hypothesis. Making the decision against or in favor of the null hypothesis reduces to determine how unlikely or likely it would be that the statistic take the value it does under the null hypothesis. There are two ways to determine this: we could consider either the so called *critical value* approach or the *p-value* approach. The first consists in determining whether the value taken by the statistic is more extreme than appropriate cut off values, computed given that the null hypothesis is true, on account of the error tolerance and alternative hypothesis. The second consists in computing the probability that the statistic takes a value at least as extreme as the one it takes in (a) direction(s) compatible with the alternative hypothesis, given that the null hypothesis is true.

In the next two sections, we will review the procedures behind each of these two approaches. To make our review concrete, we will consider the average grade μ of the students in Example 1175. We will conduct the following tests on the average grade μ each of which considers a different alternative hypothesis

$$\begin{array}{cc} \text{null hypothesis} & \text{alternative hypothesis} \\ H_0 : \mu = \mu_0 & \begin{cases} H_1 : \mu < \mu_0 \\ H_1 : \mu > \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases} , \end{array}$$

where $\mu_0 \equiv 22/30$. The first test makes sense if we are interested in concluding that the average grade is lower than μ_0, which is beyond the students' pessimistic estimate. The second test makes sense if we are interested in concluding that the average grade higher than μ_0, which is the professors' claim. The third test makes sense if we are interested in concluding that the average grade is just different than the null value μ_0. We will first consider the critical value approach and Then, the p-value approach. The procedures presented here for both approaches easily extend to tests of hypothesis on any other parameter characterizing the distribution of a population.

17.1 Rejection Region Approach

As mentioned above, the *rejection region* approach involves determining how unlikely or likely the null hypothesis is true by determining whether or not the value taken by the test statistic on the sample data collected is more extreme than some appropriate cut off value(s), known as *critical value(s)*, which are computed given that the null hypothesis is true, on account of the error tolerance and the alternative hypothesis. This entails comparing the value taken by the test statistic with the critical value(s). If the value of the test statistic is more extreme than the critical value(s) in the direction compatible with the alternative hypothesis, that is the value of the test statistic falls inside the rejection region, Then, the null hypothesis is rejected in favor of the alternative hypothesis. If the value of the test statistic is less extreme than the critical value(s), that is the valu of the test statistic falls outside the rejection region, Then, the null hypothesis is not rejected.

Specifically, the four steps involved in using the rejection region approach are:

1. identifying the parameter θ of interest and its domain Θ;

2. making the *null hypothesis*, in our case $H_0 : \theta = \theta_0$, where $\theta_0 \in \Theta$;

3. making a clear *alternative hypothesis* which contradicts the statistical hypothesis, in our case $H_1 : \theta < \theta_0$ or $H_1 : \theta > \theta_0$ or $H_1 : \theta \neq \theta_0$;

4. setting the error tolerance or *significance level* of the statistical test, that is the probability of committing an error of the I type;

5. introducing an appropriate *test statistics* $\theta_n : \Omega \to \mathbb{R}$ to deal with the null hypothesis;

6. determining the *critical value(s)* and the *rejection region* of the statistic considered, given that that the null hypothesis is true, on account of the significance level of the test and the alternative hypothesis;

7. computing the value of the test statistic on the sample data collected, given that the null hypothesis is true;

8. comparing the value of the test statistic to the critical value(s): if the value of the test statistic is more extreme than the critical value(s) in a direction compatible with the alternative hypothesis, to say the value of the test statistic falls in the rejection region, the null hypothesis can be rejected in favor of the alternative hypothesis; if the value of the test statistic is less extreme than the critical value(s), to say the value of the test statistic does not fall in the rejection region, the null hypothesis cannot be rejected.

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space representing a population and let X be a real random variable on Ω, representing a population in dependence on an unknown parameter $\theta \in \Theta \subseteq \mathbb{R}$. Let $n \in \mathbb{N}$ and let $X_1, \ldots, X_n$ be a simple random sample of size n drawn from X. In the end let $\hat{\theta}_n$ be a statistic on $X_1, \ldots, X_n$ to estimate the true value of the parameter θ.

Definition 1187 *We call the* critical value *of* $\hat{\theta}_n : \Omega \to \mathbb{R}$ *with respect to a* significance level $\alpha \in (0,1)$ *in the direction compatible with the alternative hypothesis* $H_1 : \theta < \theta_0$ *[resp.* $H_1 : \theta > \theta_0$*] the minimum [resp. maximum]* α*-quantile [resp.* $(1-\alpha)$*-quantile] of the statistic* $\hat{\theta}_n$*, given that the null hypothesis is true, that is the real number* x_α^- *[resp.* x_α^+*] such that*

$$x_\alpha^- \equiv \min_{x \in \mathbb{R}} \left\{ \mathbf{P}\left(\hat{\theta}_n \leq x \mid \theta = \theta_0\right) \geq \alpha \quad \textit{and} \quad \mathbf{P}\left(\hat{\theta}_n \geq x \mid \theta = \theta_0\right) \geq 1 - \alpha \right\} \tag{17.27}$$

$$\textit{[resp. } x_\alpha^+ \equiv \max_{x \in \mathbb{R}} \left\{ \mathbf{P}\left(\hat{\theta}_n \leq x \mid \theta = \theta_0\right) \geq 1 - \alpha \quad \textit{and} \quad \mathbf{P}\left(\hat{\theta}_n \geq x \mid \theta = \theta_0\right) \geq \alpha \right\} \textit{]} \tag{17.28}$$

(see Definition 1144 and Remark 1145). We call the critical values *of* $\hat{\theta}_n : \Omega \to \mathbb{R}$ *with respect to a* significance level $\alpha \in (0,1)$ *in the direction compatible with the alternative hypothesis* $H_1 : \theta \neq \theta_0$ *the real numbers* $x_{\alpha/2}^-$ *and* $x_{\alpha/2}^+$ *given by Equations (17.27) and (17.28) referred to the significance level* $\alpha/2$*. In what follows, the real number* x_α^- *[resp.* x_α^+*] will be also referred to as the* lower *[resp.* upper*] tail* α-critical value*. The real numbers* $x_{\alpha/2}^-$ *and* $x_{\alpha/2}^+$ *will be also referred to as the* two tails α-critical values*.*

Remark 1188 *Assume the distribution function $F_{\hat{\theta}_n} : \mathbb{R} \to \mathbb{R}$ of the test statistic $\hat{\theta}_n$, given that the null hypothesis is true, is strictly increasing and continuous. Then, x_α^- [resp. x_α^+] is the unique solution of the equations*

$$F_{\hat{\theta}_n}(x) = \alpha \quad [F_{\hat{\theta}_n}(x) = 1 - \alpha],$$

for every $\alpha \in (0,1)$.

Definition 1189 *We call the* rejection region *of a test statistic $\hat{\theta}_n : \Omega \to \mathbb{R}$ with respect to a* significance level *α in the direction of the alternative hypothesis $H_1 : \theta < \theta_0$ [resp. $H_1 : \theta > \theta_0$] the subset of the real line*

$$\left(-\infty, x_\alpha^-\right) \quad [\text{resp. } \left(x_\alpha^+, +\infty\right)],$$

where x_α^- [resp. x_α^+] is the lower [resp. upper] tail α-critical value introduced in Definition 1187, for every $\alpha \in (0,1)$. We call the rejection region *of a test statistic $\hat{\theta}_n : \Omega \to \mathbb{R}$ with respect to a* significance level *α in the direction of the alternative hypothesis $H_1 : \theta \neq \theta_0$ the subset of the real line*

$$\left(-\infty, x_{\alpha/2}^-\right) \cup \left(x_{\alpha/2}^+, +\infty\right),$$

where $x_{\alpha/2}^-$ and $x_{\alpha/2}^+$ are the two tails α-critical values also introduced in Definition 1187, for every $\alpha \in (0,1)$. In case the rejection region has the form $(-\infty, x_\alpha^-)$ [resp. $(x_\alpha^+, +\infty)$, resp. $\left(-\infty, x_{\alpha/2}^-\right) \cup \left(x_{\alpha/2}^+, +\infty\right)$] it is said to be lower tailed *[resp.* upper tailed*, resp.* two-tailed*].*

Remark 1190 *According to the critical value approach, the null hypothesis will be rejected if and only if the realization ot the test statistic on the sample data collected falls in the rejection region.*

Remark 1191 *Suppose that a null and an alternative hypothesis are made, a statistic is chosen and the sample size n is fixed. Then, decreasing [resp. increasing] the probability of committing a type I error α via the decrease of the width of the rejection region increases [resp. decreases] the probability of making a type II error β, for any value of the parameter θ consistent with H_1. A rejection reason has Then, to be chosen with an efficient compromise between α and β.*

In light of this, most statistical practitioners choose to specify the largest value of the significance level α that can be tolerated and find a rejection region in terms of specified value of α. The corresponding values of β will be the smallest possible subject to the bound on α. However, the tolerance on the value of α depends on the severity of a type I error. The more severe this error, the smaller has to be the significance level.

Example 1192 *With reference to Example 1175, assume that the random variable X expressing the mark achieved by a student in the exam is binomially distributed with parameters $m = 30$ and p. Apply the critical value approach to test the null hypothesis that students' rumors about the average grade $\mu_0 = 20$ achieved at the exam are right at the significance level $\alpha = 0.05$.*

Discussion. The assumption that the random variable X expressing the mark achieved by a student in the exam is binomially distributed with parameters $m = 30$ and p, that is $X \sim Bin(m, p)$, and that $\mu_0 = 20$, leads to the equation

$$\mu_0 = \mathbf{E}[X] = mp = 20$$

yielding

$$p = \frac{20}{30} = \frac{2}{3}.$$

About the validity of the model, consider that the probability that a student does not pass the exam is given by

$$\sum_{j=0}^{17} \binom{30}{j} \cdot \left(\frac{2}{3}\right)^{j} \cdot \left(1 - \frac{2}{3}\right)^{30-j} \approx 0.166 = 16.6\%.$$

In addition, note that the probability that a student passes the exams with a mark not higher than 24 is given by

$$\sum_{j=18}^{24} \binom{30}{j} \cdot \left(\frac{2}{3}\right)^{j} \cdot \left(1 - \frac{2}{3}\right)^{30-j} \approx 0.799 = 79.9\%$$

and the probability that a student passes the exams with a mark higher than 24 is given by

$$\sum_{j=25}^{30} \binom{30}{j} \cdot \left(\frac{2}{3}\right)^{j} \cdot \left(1 - \frac{2}{3}\right)^{30-j} \approx 0.035 = 3.5\%.$$

Hence, despite the model is quite rough, yet it might be useful as a first approximation. A simple random sample of nine students is selected end the following evaluation on the marks x_k, for $k = 1, \ldots, 9$, achieved by the students is computed

$$\textstyle\sum_{k=1}^{9} x_k = 197.$$

Unfortunately, this assumption makes it difficult to deal with a statistic on a simple random sample drawn from X. However, the students' rumors about the average grade achieved at the exam imply that $p = 20/30$. Thus, the rule of thumb

$$mp \geq 5 \quad \text{and} \quad m\left(1 - p\right) \geq 5$$

under which we can approximate a binomial distribution with a normal distribution is largely fullfilled. Therefore, we consider a normal approximation of X which is a normally distributed random variable with mean $\mu = \mathbf{E}\left[X\right] = mp$ and variance $\sigma^2 = \mathbf{D}^2\left[X\right] = mp\left(1 - p\right)$. That is $X \sim N\left(\mu, \sigma^2\right)$. Now, considering a simple random sample $X_1, \ldots, X_n$ of size n drawn from X, we can think that the statistic

$$Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}},$$

where $\bar{X}_n$ is the sample mean of $X_1, \ldots, X_n$, is normally distributed with mean 0 and variance 1. In particular, assuming that the null hypothesis $H_0 : \mu = \mu_0$ is true for $\mu_0 = 20$, we have that the statistic

$$Z_0 = \frac{\bar{X}_n - \mu_0}{\sigma_0/\sqrt{n}} = \frac{\bar{X}_n - mp_0}{\sqrt{mp_0\left(1 - p_0\right)/n}},$$

where $p_0 \equiv \mu_0/m$, is normally distributed with mean 0 and variance 1. Now, with reference to the alternative hypothesis $H_1 : \mu < \mu_0$, for the probability of type I error $\alpha = 0.05$, we obtain

the rejection region $(-\infty, x_\alpha^-)$, where $x_\alpha^- = -1.645$. On the other hand, the value taken by Z_0 on the sample data collected is

$$Z_0(\omega) = \frac{\frac{1}{9}\sum_{k=1}^{9} x_k - 30 \cdot \frac{20}{30}}{\sqrt{30 \cdot \frac{20}{30}\left(1 - \frac{20}{30}\right)/9}} = \frac{\frac{1}{9} \cdot 197 - 20}{\sqrt{20 \cdot \frac{1}{3} \cdot \frac{1}{9}}} = 2.195$$

which does not fall in the rejection region. This means that we cannot reject the null hypothesis $H_0 : \mu = \mu_0$ in favor of the alternative hypothesis $H_1 : \mu < \mu_0$, with the significance level $\alpha = 0.05$. In other words, there is no reason for the students to be more pessimistic than they already are. With reference to the alternative hypothesis $H_1 : \mu > \mu_0$, for the probability of type I error $\alpha = 0.05$, we obtain the rejection region $(z_\alpha^+, +\infty)$, where $z_\alpha^+ = 1.645$. In this case, the observed value $Z_0(\omega)$ of our statistic falls far inside the rejection region. This means that we can reject the null hypothesis $H_0 : \mu = \mu_0$ in favor of the alternative hypothesis $H_1 : \mu > \mu_0$, with the significance level $\alpha = 0.05$. In other words, the professors have some reason to push the students to be more optimistic. In the end, with reference to the alternative hypothesis $H_1 : \mu \neq \mu_0$, still for a the probability of type I error $\alpha = 0.05$, we obtain the rejection region $\left(-\infty, z_{\alpha/2}^-\right) \cup \left(z_{\alpha/2}^+, +\infty\right)$, where $z_{\alpha/2}^- = -1.960$ and $z_{\alpha/2}^+ = 1.960$. Again, the value $Z_0(\omega)$ taken by our statistic falls far inside the rejection region. This means that in any case the students' rumors are wrong at the significance level $\alpha = 0.05$. It may be interesting to compute the probability of committing a type II error, that is the probability of not rejecting the assumption $H_0 : \mu = \mu_0$ when it is false. With reference to the alternative hypothesis $H_1 : \mu > \mu_0$ and a significance level $\alpha = 0.05$ we fail to reject the assumption $H_0 : \mu = \mu_0$ when

$$Z_0 = \frac{\bar{X}_n - mp_0}{\sqrt{mp_0(1-p_0)/n}} \leq z_\alpha^+,$$

where $z_\alpha^+ = 1.645$. Equivalently,

$$\bar{X}_n \leq mp_0 + z_\alpha^+ \sqrt{mp_0(1-p_0)/n} = 30 \cdot \frac{20}{30} + 1.645 \cdot \sqrt{30 \cdot \frac{20}{30}\left(1 - \frac{20}{30}\right)/9} = 21.416.$$

Thus, we fail to reject H_0 when $\bar{X}_n \leq 21.416$. Note that the value $\bar{X}_n(\omega)$ taken by $\bar{X}_n$ on our sample is

$$\bar{X}_n(\omega) = 21.889.$$

On the other hand, H_0 is false when $\mu \neq \mu_0$ in the direction compatible with the alternative hypothesis. For instance, referring to $H_1 : \mu > \mu_0$, we could consider $\mu = \mu_1$ for $\mu_1 \equiv 21$, $\mu_1 \equiv 22$, $\mu_1 \equiv 23$, $\mu_1 \equiv 24$, $\mu_1 \equiv 25$, $\mu_1 \equiv 26$, $\mu_1 \equiv 27$, $\mu_1 \equiv 28$, $\mu_1 \equiv 29$, $\mu_1 \equiv 30$. Now, assuming $\mu = \mu_1$ for a mispecified μ_1 implies that the statistic

$$Z_1 \equiv \frac{\bar{X}_n - mp_1}{\sqrt{mp_1(1-p_1)/n}}$$

is normally distributed with mean 0 and variance 1. On the other hand, since we can write

$$\begin{aligned}\frac{\bar{X}_n - mp_0}{\sqrt{mp_0(1-p_0)/n}} &= \frac{\sqrt{mp_1(1-p_1)/n}}{\sqrt{mp_1(1-p_1)/n}}\frac{\bar{X}_n - mp_1 + mp_1 - mp_0}{\sqrt{mp_0(1-p_0)/n}} \\ &= \frac{\sqrt{p_1(1-p_1)}}{\sqrt{p_0(1-p_0)}}\frac{\bar{X}_n - mp_1}{\sqrt{mp_1(1-p_1)/n}} + \frac{m(p_1-p_0)}{\sqrt{mp_0(1-p_0)/n}} \\ &= \frac{\sqrt{p_1(1-p_1)}}{\sqrt{p_0(1-p_0)}}Z_1 + \frac{\sqrt{m}(p_1-p_0)}{\sqrt{p_0(1-p_0)/n}}\end{aligned}$$

we have

$$Z_0 = \frac{\sqrt{p_1(1-p_1)}}{\sqrt{p_0(1-p_0)}}Z_1 + \frac{\sqrt{m}(p_1-p_0)}{\sqrt{p_0(1-p_0)/n}}.$$

It follows

$$\begin{aligned}\beta(\mu_1) &\equiv \mathbf{P}(\text{not reject } H_0 \mid \mu = \mu_1) = \mathbf{P}\left(Z_0 \leq z_\alpha^+ \mid \mu = \mu_1\right) \\ &= \mathbf{P}\left(\frac{\sqrt{p_1(1-p_1)}}{\sqrt{p_0(1-p_0)}}Z_1 + \frac{\sqrt{m}(p_1-p_0)}{\sqrt{p_0(1-p_0)/n}} \leq z_\alpha^+ \mid \mu = \mu_1\right) \\ &= \mathbf{P}\left(Z_1 \leq \frac{\sqrt{p_0(1-p_0)}}{\sqrt{p_1(1-p_1)}}\left(z_\alpha^+ - \frac{\sqrt{m}(p_1-p_0)}{\sqrt{p_0(1-p_0)/n}}\right) \mid \mu = \mu_1\right) \\ &= \Phi\left(\frac{\sqrt{p_0(1-p_0)}}{\sqrt{p_1(1-p_1)}}\left(z_\alpha^+ - \frac{\sqrt{m}(p_1-p_0)}{\sqrt{p_0(1-p_0)/m}}\right)\right).\end{aligned}$$

We have

$$\frac{\sqrt{p_0(1-p_0)}}{\sqrt{p_1(1-p_1)}}\left(z_\alpha^+ - \frac{\sqrt{m}(p_1-p_0)}{\sqrt{p_0(1-p_0)/n}}\right) \approx \begin{cases} 0.497 & \text{for } \mu_1 = 21 \\ -0.724 & \text{for } \mu_1 = 22 \\ -2.052 & \text{for } \mu_1 = 23 \\ -3.539 & \text{for } \mu_1 = 24 \\ -5.268 & \text{for } \mu_1 = 25 \\ -7.386 & \text{for } \mu_1 = 26 \\ -10.195 & \text{for } \mu_1 = 27 \\ -14.457 & \text{for } \mu_1 = 28 \\ -23.142 & \text{for } \mu_1 = 29 \\ -\infty & \text{for } \mu_1 = 30 \end{cases}.$$

Therefore,

$$\beta(\mu_1) = \begin{cases} 0.690 & \text{for } \mu_1 = 21 \\ 0.234 & \text{for } \mu_1 = 22 \\ 2.008 \times 10^{-2} & \text{for } \mu_1 = 23 \\ 2.008 \times 10^{-4} & \text{for } \mu_1 = 24 \\ 6.896 \times 10^{-8} & \text{for } \mu_1 = 25 \\ 7.566 \times 10^{-14} & \text{for } \mu_1 = 26 \\ 1.044 \times 10^{-24} & \text{for } \mu_1 = 27 \\ 1.132 \times 10^{-47} & \text{for } \mu_1 = 28 \\ 8.750 \times 10^{-119} & \text{for } \mu_1 = 29 \\ 0 & \text{for } \mu_1 = 30 \end{cases}.$$

Note that average grades μ higher than 24 the probability of accepting H_0 when H_0 is false is very low. □

17.2 P-value approach

The *p-value* approach involves determining how unlikely or likely the null hypothesis is true by determining the probability of observing a vaule of the test statistic at least as extreme as the one acually observed, in the direction compatible with the alternative hypothesis, given that the null hypothesis is true. If the p-value is small Then, it is unlikely that the null hypothesis is true against the alternative hypothesis. If the p-value is large Then, it is likely that the null hypothesis is true. Eventually, the smaller the P-value is the more unlikely the null hypothesis is true against the alternative hypothesis.

The steps involved in using the P-value approach to conduct any hypothesis test are almost the same than in case of the critical value approach:

1. identifying the parameter θ of interest and its domain Θ;
2. making the *null hypothesis*, in our case $H_0 : \theta = \theta_0$, where $\theta_0 \in \Theta$;
3. making a clear *alternative hypothesis* which contradicts the statistical hypothesis, in our case $H_1 : \theta < \theta_0$ or $H_1 : \theta > \theta_0$ or $H_1 : \theta \neq \theta_0$;
4. introducing an appropriate *test statistics* $\theta_n : \Omega \to \mathbb{R}$ to deal with the null hypothesis;
5. computing the value of the test statistic on the sample data collected, given that that the null hypothesis is true;
6. considering the distribution of the test statistic to calculate the p-value which answer the question: given that the null hypothesis is true, what is the probability that we would have observed a more extreme test statistic, in the direction compatible with the alternative hypothesis, than we did? (Note how this question is equivalent to the question answered in criminal trials: given that the defendant is innocent, what is the chance that we would have observed such an extreme criminal evidence?)
7. comparing the p-value to the significance level α chosen in advance: if the p-value is smaller than α, Then, the null hypothesis can be rejected in favor of the alternative hypothesis, at the chosen significance level; if the p-value is greater than α, the null hypothesis cannot be rejected against the alternative hypothesis, at the chosen significance level.

Note that the p-value approach always leads to the same conclusion of the rejection region approach: the null hypothesis is rejected in favor of [resp. not rejected against] the alternative hypothesis in light of the observed value of the test statistic at a given significance level α via the p-value approach, if and only if the observed value of the test statistic falls [does not fall] in the rejection region characterized by the significance level α.

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space representing a population and let X be a real random variable on Ω, which represents a characteristic of the elements in the population with distribution depending on an unknown parameter $\theta \in \Theta \subseteq \mathbb{R}$. Let $n \in \mathbb{N}$ and let $X_1, \ldots, X_n$ be a simple random sample of size n drawn from X. In the end, let $\hat{\theta}_n : \Omega \to \mathbb{R}$ a point estimator of θ and let $\hat{\theta}_n(\omega) \equiv \theta(X_1(\omega), \ldots, X_n(\omega))$ the point estimate of θ on the sample.

Definition 1193 *Given the null hypothesis $H_0 : \theta = \theta_0$, we call the* p-value *of $\hat{\theta}_n$ in light of the observed value $\hat{\theta}_n(\omega)$ the probability*

$$\mathbf{P}\left(\hat{\theta}_n < \hat{\theta}_n(\omega) \mid \theta = \theta_0\right) \quad or \quad \mathbf{P}\left(\hat{\theta}_n > \hat{\theta}_n(\omega) \mid \theta = \theta_0\right) \tag{17.29}$$

or

$$\mathbf{P}\left(\hat{\theta}_n < -\left|\hat{\theta}_n(\omega)\right| \mid \theta = \theta_0\right) + \mathbf{P}\left(\theta_n > \left|\hat{\theta}_n(\omega)\right| \mid \theta = \theta_0\right), \tag{17.30}$$

according to whether we consider the alternative hypothesis $H_1 : \theta < \theta_0$ *or* $H_1 : \theta > \theta_0$ *or* $H_1 : \theta \neq \theta_0$.

Definition 1194 *Given the null hypothesis* $H_0 : \theta = \theta_0$ *and a significance level* α*, typically chosen among* 0.10, 0.05, *and* 0.01, *we say that* the null hypothesis is rejected at the significance level α in favor of the alternative hypothesis $H_1 : \theta < \theta_0$ *or* $H_1 : \theta > \theta_0$ *or* $H_1 : \theta \neq \theta_0$ *according to whether we have*

$$\mathbf{P}\left(\hat{\theta}_n < \hat{\theta}_n(\omega) \mid \theta = \theta_0\right) \leq \alpha \quad or \quad \mathbf{P}\left(\hat{\theta}_n < -\left|\hat{\theta}_n(\omega)\right| \mid \theta = \theta_0\right) \leq \alpha \tag{17.31}$$

or

$$\mathbf{P}\left(\hat{\theta}_n < -\left|\hat{\theta}_n(\omega)\right| \mid \theta = \theta_0\right) + \mathbf{P}\left(\theta_n > \left|\hat{\theta}_n(\omega)\right| \mid \theta = \theta_0\right) \leq \alpha. \tag{17.32}$$

Otherwise, we say that the null hypothesis is not rejected at the significance level α against the alternative hypothesis.

Example 1195 *With reference to Example 1192, apply the critical value approach to test the null hypothesis that students's rumors about the average grade* $\mu_0 = 20$ *achieved at the exam are right at the significance level* $\alpha = 0.05$.

Discussion. As discussed in Example 1192, assuming that the null hypothesis $H_0 : \mu = \mu_0$ is true for $\mu_0 = 20$, we have that the statistic

$$Z_0 = \frac{\bar{X}_n - \mu_0}{\sigma_0/\sqrt{n}} = \frac{\bar{X}_n - mp_0}{\sqrt{mp_0(1-p_0)/n}},$$

where $p_0 = \mu_0/m$, is normally distributed with mean 0 and variance 1. In addition, we know that the observed realization of Z_0 on the sample data collected is

$$\hat{z}_0 = 2.195.$$

Now, in a direction compatible with the alternative hypothesis $H_1 : \mu < \mu_0$, we have

$$\mathbf{P}(Z_0 < \hat{z}_0 \mid \mu = \mu_0) = \mathbf{P}(Z_0 \leq \hat{z}_0 \mid Z_0 \sim N(0,1)) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\hat{z}_0} e^{-\frac{x^2}{2}} dx \simeq 0.986,$$

which implies that the null hypothesis cannot be rejected in favor of the alternative hypothesis $H_1 : \mu < \mu_0$ at no significance level among the typical ones. On the other hand, in a direction compatible with the alternative hypothesis $H_1 : \mu < \mu_0$, we have

$$\mathbf{P}(Z_0 > \hat{z}_0 \mid \mu = \mu_0) = 1 - \mathbf{P}(Z_0 \leq \hat{z}_0 \mid Z_0 \sim N(0,1)) = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\hat{z}_0} e^{-\frac{x^2}{2}} dx \simeq 0.014,$$

which implies that the null hypothesis can be rejected in favor of the alternative hypothesis $H_1 : \mu > \mu_0$ at any significance level among the typical ones. In the end, in a direction compatible with the alternative hypothesis $H_1 : \mu \neq \mu_0$, we have

$$\mathbf{P}(Z_0 < -\hat{z}_0 \mid \mu = \mu_0) + \mathbf{P}(Z_0 > \hat{z}_0 \mid \mu = \mu_0) = 2\mathbf{P}(Z_0 > \hat{z}_0 \mid \mu = \mu_0) \simeq 0.028,$$

which again implies that the null hypothesis can be rejected in favor of the alternative hypothesis $H_1 : \mu \neq \mu_0$ at any significance level among the typical ones. □

17.3 Hypothesis Testing about a Population Mean

17.3.1 Any Sample from a Normal Population with Known Variance

Let X a real random variable representing a population. We assume that X is normally distributed with unknown mean μ, but known variance σ^2. In symbols $X \sim N(\mu, \sigma^2)$. Although the assumption that the variance is known when variance is not is quite unrealistic, this simple case costitutes a useful starting point towards more general cases. We make the null hypothesis in the form $H_0 : \mu = \mu_0$, where μ_0 is a suitable real number, and the alternative hypothesis in each of the forms $H_1 : \mu < \mu_0$, $H_1 : \mu > \mu_0$, and $H_1 : \mu \neq \mu_0$. Let $X_1, \ldots, X_n$ be a simple random sample of size n drawn from X and let $\bar{X}_n$ be the sample mean of X. We know that $\bar{X}_n$ is normally distributed with mean $\mu_{\bar{X}_n} = \mu$ and variance $\sigma^2_{\bar{X}_n} = \sigma^2/n$. Hence, standardizing $\bar{X}_n$, we obtain a normally distributed random variable

$$Z \equiv \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

with mean 0 and variance 1. In symbols, $Z \sim N(0, 1)$. We choose the random variable Z as our test statistic since the the critical values of the distribution of Z are well known. If we chose $\bar{X}_n$ as test statistic, Then, to determine the critical values of $\bar{X}_n$ we should standardize it!

Assuming the null hypothesis true, that is $\mu = \mu_0$, Then, also the random variable

$$Z_0 \equiv \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}}$$

is normally distributed with with mean 0 and variance 1. Therefore, choosing a significance level $\alpha \in (0, 1)$ we can determine easily the critical value(s) of Z_0 corresponding to each of the alternative hypotheses considered. We report the critical value(s) and the rejection region for the cases $\alpha \equiv 0.10$, 0.05, and 0.01 in Table (17.33).

$$
\begin{array}{lll}
\text{null hypothesis} & H_0 : \mu = \mu_0 & \\
\text{alternative hypothesis} & \text{significance level} & \text{critical value(s)} \\
H_1 : \mu < \mu_0 & \alpha \equiv 0.10 & z_\alpha^- = -1.282 \\
& \alpha \equiv 0.05 & z_\alpha^- = -1.645 \\
& \alpha \equiv 0.01 & z_\alpha^- = -2.326 \\
H_1 : \mu > \mu_0 & \alpha \equiv 0.10 & z_\alpha^+ = 1.282 \\
& \alpha \equiv 0.05 & z_\alpha^+ = 1.645 \\
& \alpha \equiv 0.01 & z_\alpha^+ = 2.326 \\
H_1 : \mu \neq \mu_0 & \alpha \equiv 0.10 & z_{\alpha/2}^- = -1.645,\ z_{\alpha/2}^+ = 1.645 \\
& \alpha \equiv 0.05 & z_{\alpha/2}^- = -1.960,\ z_{\alpha/2}^+ = 1.960 \\
& \alpha \equiv 0.01 & z_{\alpha/2}^- = -2.257,\ z_{\alpha/2}^+ = 2.257
\end{array}
$$

$$
\begin{array}{ll}
\text{rejection region} & \text{probability of type I error} \\
\left(-\infty, z_\alpha^-\right) & \alpha = \mathbf{P}\left(Z_0 < z_\alpha^-\right) \\
\left(z_\alpha^+, +\infty\right) & \alpha = \mathbf{P}\left(Z_0 > z_\alpha^+\right) \\
\left(-\infty, z_{\alpha/2}^-\right) \cup \left(z_{\alpha/2}^+, +\infty\right) & \begin{array}{l} \alpha/2 = \mathbf{P}\left(Z_0 < z_{\alpha/2}^-\right) \\ \alpha/2 = \mathbf{P}\left(Z_0 > z_{\alpha/2}^+\right) \end{array}
\end{array}
$$

(17.33)

Recall that, by virtue of the symmetry of the normal distribution, we have

$$z_{\alpha}^{-} = -z_{\alpha}^{+} \quad \text{and} \quad z_{\alpha/2}^{-} = -z_{\alpha/2}^{+}.$$

More importantly, note that

$$\mathbf{P}\left(\text{reject } H_0 \mid H_0 \text{ is true}\right) = \begin{array}{ll} \mathbf{P}\left(Z_0 < z_{\alpha}^{-}\right) & \text{if } H_1 : \mu < \mu_0 \\ \mathbf{P}\left(Z_0 > z_{\alpha}^{+}\right) & \text{if } H_1 : \mu > \mu_0 \\ \mathbf{P}\left(Z_0 < z_{\alpha/2}^{-}\right) + \mathbf{P}\left(Z_0 > z_{\alpha/2}^{+}\right) & \text{if } H_1 : \mu \neq \mu_0 \end{array}.$$

For instance, referring to the alternative hypothesis $H_1 : \mu > \mu_0$, this means that under the null hypothesis $H_0 : \mu = \mu_0$ the probability that the statistic Z_0 takes a value larger than z_{α}^{+} is equal to α, which is what matters according to the alternative hypothesis. Once more note that when we observe a realization $Z(\omega) > z_{\alpha}^{+}$ we cannot speak in terms of probability. Probability evaluates the possibility of occurrence of events before they occur! However, analogously to the case of the confidence intervals, we speak in terms of significance level. More precisely, we say that we reject the null hypothesis with a significance level α. On the contrary, if we observe a realization $Z(\omega) \leq z_{\alpha}^{+}$ we cannot reject the null hypothesis with the same significance level α.

To determine the probability β of committing a type II error, which is

$$\beta \equiv \mathbf{P}\left(\text{not reject } H_0 \mid H_0 \text{ is false}\right),$$

we observe first that declaring H_0 false has different implications according to the alternative hypothesis under consideration. For instance, declaring H_0 false in a direction compatible with the alternative hypothesis $H_1 : \mu > \mu_0$ means that true value of the parameter μ has to be a value $\mu_1 > \mu_0$. However, since we do not know the true value of the parameter μ, we need to assume that the hypothesis

$$H_0^{(1)} : \mu = \mu_1$$

might be true for each $\mu_1 > \mu_0$. This implies that we will determine β as a function μ_1. To this goal, it is important to notice that assuming $H_0^{(1)}$ true, for each $\mu_1 > \mu_0$, leads to consider the statistic

$$Z_1 \equiv \frac{\bar{X}_n - \mu_1}{\sigma/\sqrt{n}}$$

as normally distributed with mean 0 and variance 1. On the other hand, we do not reject H_0 when we observe a realization of $Z_0 \leq z_{\alpha}^{+}$. In light of the above arguments, we can write

$$\begin{aligned} \beta(\mu_1) &= \mathbf{P}\left(\text{not reject } H_0 \mid H_0^{(1)}\right) \\ &= \mathbf{P}\left(Z_0 \leq z_{\alpha}^{+} \mid \mu = \mu_1\right) \\ &= \mathbf{P}\left(\frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \leq z_{\alpha}^{+} \mid \mu = \mu_1\right) \\ &= \mathbf{P}\left(\bar{X}_n \leq \mu_0 + z_{\alpha}^{+}\sigma/\sqrt{n} \mid \mu = \mu_1\right) \\ &= \mathbf{P}\left(\frac{\bar{X}_n - \mu_1}{\sigma/\sqrt{n}} \leq \frac{\mu_0 - \mu_1 + z_{\alpha}^{+}\sigma/\sqrt{n}}{\sigma/\sqrt{n}} \mid \mu = \mu_1\right) \\ &= \mathbf{P}\left(Z_1 \leq z_{\alpha}^{+} - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right) \\ &= \Phi\left(z_{\alpha}^{+} - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right). \end{aligned}$$

which yields a close formula to compute the probability of committing a type II error given a significance level α in a direction compatible with the alternative hypothesis $H_1 : \mu > \mu_0$.

In terms of p-value the argument is rather simple. If we assume that H_0 is true, that is $\mu = \mu_0$, Then, the random variable

$$Z_0 \equiv \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}}$$

is normally distributed with with mean 0 and variance 1. Therefore, for any realization $Z_0(\omega)$ of Z_0 it is easy to compute the probability that Z_0 takes a value larger than the realized one. Then, fixed any significance level $\alpha \in (0, 1)$ we reject the null hypothesis whether

$$\mathbf{P}(Z_0 > Z_0(\omega)) < \alpha$$

or not.

Similar arguments apply when we consider the cases related to the alternative hypothesis $H_1 : \mu < \mu_0$ and $H_1 : \mu \neq \mu_0$. In particular, also in these cases we can determine close formulas for the type II error probabilities. We summarize these formulas in Table (17.34).

null hypothesis $H_0 : \mu = \mu_0$

alternative hypothesis	significance level	critical value(s)	rejection region	probability of type II error
$H_1 : \mu < \mu_0$	$\alpha \equiv 0.10$	$z_\alpha^- = -1.282$	$(-\infty, z_\alpha^-)$	$\beta(\mu_1) = 1 - \Phi\left(z_\alpha^- + \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}}\right)$
	$\alpha \equiv 0.05$	$z_\alpha^- = -1.645$		
	$\alpha \equiv 0.01$	$z_\alpha^- = -2.326$		
$H_1 : \mu > \mu_0$	$\alpha \equiv 0.10$	$z_\alpha^+ = 1.282$	$(z_\alpha^+, +\infty)$	$\beta(\mu_1) = \Phi\left(z_\alpha^+ - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right)$
	$\alpha \equiv 0.05$	$z_\alpha^+ = 1.645$		
	$\alpha \equiv 0.01$	$z_\alpha^+ = 2.326$		
$H_1 : \mu \neq \mu_0$	$\alpha \equiv 0.10$	$z_{\alpha/2}^- = -1.645,\ z_{\alpha/2}^+ = 1.645$	$\left(-\infty, z_{\alpha/2}^-\right) \cup \left(z_{\alpha/2}^+, +\infty\right)$	$\beta(\mu_1) = \Phi\left(z_{\alpha/2}^+ - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right) - \Phi\left(z_{\alpha/2}^- + \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}}\right)$
	$\alpha \equiv 0.05$	$z_{\alpha/2}^- = -1.960,\ z_{\alpha/2}^+ = 1.960$		
	$\alpha \equiv 0.01$	$z_{\alpha/2}^- = -2.257,\ z_{\alpha/2}^+ = 2.257$		

(17.24)

From the formulas yielding the probability of a type II error appears the possibility of controlling such a probability by a suitable choice of the size n of the data sample. To this goal, setting $\beta \equiv \beta(\mu_1)$, determine z_β such that

$$\beta = \Phi(-z_\beta) = 1 - \Phi(z_\beta).$$

Then, we can write

$$\Phi(-z_\beta) = \Phi\left(z_\alpha^+ - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right),$$

which implies

$$-z_\beta = z_\alpha^+ - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}.$$

From the latter, it follows

$$n = \left(\frac{\sigma(z_\beta + z_\alpha^+)}{\mu_1 - \mu_0}\right)^2. \tag{17.35}$$

Similarly, writing

$$1 - \Phi(z_\beta) = 1 - \Phi\left(z_\alpha^- + \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}}\right),$$

we obtain

$$z_\beta = z_\alpha^- + \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}}.$$

It follows

$$n = \left(\frac{\sigma(z_\beta - z_\alpha^-)}{\mu_0 - \mu_1}\right)^2. \tag{17.36}$$

Note that, since $z_\alpha^- = z_\alpha^+$, Formula (17.35) for an upper tailed test coincides with formula (17.36) for a lower tailed test. In the end, it is possible to prove that the formula for a two-tailed test is given by

$$n = \left(\frac{\sigma\left(z_\beta + z_{\alpha/2}^+\right)}{\mu_0 - \mu_1}\right)^2. \tag{17.37}$$

17.3.2 Any Sample from a Normal Population with Unknown Variance

Let X a real random variable representing a population. We assume that X is (at least approximately) normally distributed with unknown mean μ and unknown variance σ^2. In symbols $X \sim N(\mu, \sigma^2)$. Still, we make the null hypothesis in the form $H_0 : \mu = \mu_0$, where μ_0 is a suitable real number, and we consider the alternative hypothesis in each of the forms $H_1 : \mu > \mu_0$, $H_1 : \mu < \mu_0$, and $H_1 : \mu \neq \mu_0$. Let $X_1, \ldots, X_n$ be a simple sample of size n drawn from X and let $\bar{X}_n$ and S_n be the sample mean and sample standard deviation of X. We know that standardizing $\bar{X}_n$ via the unbiased sample standard deviation S_n the random variable

$$T_{n-1} \equiv \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$$

has a Student t distribution with $n-1$ degree of freedom. If we assume that H_0 is true, that is $\mu = \mu_0$, Then, also the random variable

$$T_{n-1}^{(0)} \equiv \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}}$$

has a Student t distribution with $n-1$ degree of freedom. Therefore, choosing a significance level $\alpha \in (0,1)$ we can determine easily the critical value(s) of $T_{n-1}^{(0)}$ corresponding to each of the alternative hypotheses considered. We report the critical value(s) and the rejection region for the cases $\alpha \equiv 0.10$, 0.05, 0.01, and $n = 4, 5, 6, 7$ in Table (17.38).

null hypothesis $H_0 : \mu = \mu_0$

signific. lev.	critical value(s) $n = 4, 5, 6, 7$	rejection region	prob. of type I error
$\alpha \equiv 0.10$	$t^-_{3,\alpha} = -1.638$, $t^-_{4,\alpha} = -1.533$, $t^-_{5,\alpha} = -1.476$, $t^-_{6,\alpha} = -1.440$,	$\left(-\infty, t^-_{n-1,\alpha}\right)$	$\alpha = \mathbf{P}\left(Z_0 < t^-_{n-1,\alpha}\right)$
$\alpha \equiv 0.05$	$t^-_{3,\alpha} = -2.353$, $t^-_{4,\alpha} = -2.132$, $t^-_{5,\alpha} = -2.015$, $t^-_{6,\alpha} = -1.943$,		
$\alpha \equiv 0.01$	$t^-_{3,\alpha} = -4.541$, $t^-_{4,\alpha} = -3.747$, $t^-_{5,\alpha} = -3.365$, $t^-_{6,\alpha} = -3.143$,		
$\alpha \equiv 0.10$	$t^+_{3,\alpha} = 1.638$, $t^+_{4,\alpha} = 1.533$, $t^+_{5,\alpha} = 1.476$, $t^+_{6,\alpha} = 1.440$,	$\left(t^+_{n-1,\alpha}, +\infty\right)$	$\alpha = \mathbf{P}\left(Z_0 > t^+_{n-1,\alpha}\right)$
$\alpha \equiv 0.05$	$t^+_{3,\alpha} = 2.353$, $t^+_{4,\alpha} = 2.132$, $t^+_{5,\alpha} = 2.015$, $t^+_{6,\alpha} = 1.943$,		
$\alpha \equiv 0.01$	$t^+_{3,\alpha} = 4.541$, $t^+_{4,\alpha} = 3.747$, $t^+_{5,\alpha} = 3.365$, $t^+_{6,\alpha} = 3.143$,		
$\alpha \equiv 0.10$	$t^-_{3,\alpha/2} = -2.353$, $t^-_{4,\alpha/2} = -2.132$, $t^-_{5,\alpha/2} = -2.015$, $t^-_{6,\alpha/2} = -1.943$, $t^+_{3,\alpha/2} = 2.353$, $t^+_{4,\alpha/2} = 2.132$, $t^+_{5,\alpha/2} = 2.015$, $t^+_{6,\alpha/2} = 1.943$,	$\left(-\infty, t^-_{n-1,\alpha/2}\right) \cup \left(t^+_{n-1,\alpha/2}, +\infty\right)$	$\alpha/2 = \mathbf{P}\left(Z_0 < t^-_{n-1,\alpha/2}\right)$
$\alpha \equiv 0.05$	$t^-_{3,\alpha/2} = -3.182$, $t^-_{4,\alpha/2} = -2.776$, $t^-_{5,\alpha/2} = -2.571$, $t^-_{6,\alpha/2} = -2.447$, $t^+_{3,\alpha/2} = 3.182$, $t^+_{4,\alpha/2} = 2.776$, $t^+_{5,\alpha/2} = 2.571$, $t^+_{6,\alpha/2} = 2.447$,		$\alpha/2 = \mathbf{P}\left(Z_0 > t^-_{n-1,\alpha/2}\right)$
$\alpha \equiv 0.01$	$t^-_{3,\alpha/2} = -5.841$, $t^-_{4,\alpha/2} = -4.604$, $t^-_{5,\alpha/2} = -4.032$, $t^-_{6,\alpha/2} = -3.707$, $t^+_{3,\alpha/2} = 5.841$, $t^+_{4,\alpha/2} = 4.604$, $t^+_{5,\alpha/2} = 4.032$, $t^+_{6,\alpha/2} = 3.707$,		

Recall that, by virtue of the symmetry of the Student t distribution, we have

$$t^{-}_{n-1,\alpha} = -t^{+}_{n-1,\alpha} \quad \text{and} \quad t^{-}_{n-1,\alpha/2} = -t^{+}_{n-1,\alpha/2}.$$

Note also that the rejection regions in Table (17.38) differs from those in Table (17.33) only in that the t critical values $t^{\mp}_{n-1,\alpha}$ or $t^{\mp}_{n-1,\alpha/2}$ replace the z critival values $z^{\mp}_{\alpha}$ or $z^{\mp}_{\alpha/2}$, respectively.

To determine the probability β of committing a type II error, the same argument exploited in Subsection 17.3.1 applies. For instance, with reference to the alternative hypothesis $H_1 : \mu > \mu_0$, it follows that we will determine β as a function $\mu_1 > \mu_0$, assuming that the hypothesis

$$H_0^{(1)} : \mu = \mu_1$$

might be true for each $\mu_1 > \mu_0$. However, in this case the computation of

$$\beta(\mu_1) = \mathbf{P}\left(\text{not reject } H_0 \mid H_0^{(1)}\right) = \mathbf{P}\left(T^{(0)}_{n-1} \leq t^{+}_{n-1,\alpha} \mid \mu = \mu_1\right)$$

can be accomplished only with numerical methods. Moreover, it depends on n and on the ratio

$$d \equiv \frac{\mu_1 - \mu_0}{\sigma}.$$

Hence, the probability of a type II error is eventually a function of n and d

$$\mathbf{P}\left(T^{(0)}_{n-1} \leq t^{+}_{n-1,\alpha} \mid \mu = \mu_1\right) = \beta(n, d).$$

Given the typical values of α, this computation can be commonly found in statistical formularies for several values of n and on varying of d in large intervals. For each value of n, a graph is available which allows to know the value of $\beta(n, d)$ on the vertical axis corresponding to the value of d on the horizontal axis. The caveat is that to compute d with reference to the values of the parameter value μ_1 of our concern, we need to guess the standard deviation σ of the distribution of X. For each fixed μ_1, the larger σ is, the larger β follows. Note that these graphs are also used when, for a given α, some μ_1 of our concern, and a guessed σ, we are interested in determining the sample size n to keep β below a given threeshold. To this goal, we select n in such a way that $n - 1$ corresponds to the closest graph with respect to which the point (d, β) lies above.

Similar arguments hold true with reference to the alternative hypothesis $H_1 : \mu < \mu_0$ and $H_1 : \mu \neq \mu_0$.

17.3.3 Large Sample from a Non-Normal Population with Unknown Variance

Let X be a real random variable representing a population. We assume that the distribution of X is unknown as well as its mean μ and variance σ^2. Still, we make the null hypothesis in the form $H_0 : \mu = \mu_0$, where μ_0 is a suitable real number, and we consider the alternative hypothesis in each of the forms $H_1 : \mu > \mu_0$, $H_1 : \mu < \mu_0$, and $H_1 : \mu \neq \mu_0$. Let $X_1, \ldots, X_n$ be a simple sample of size n drawn from X and let $\bar{X}_n$ and S_n be the sample mean and sample standard deviation of X. We know that if n is large[1] Then, the statistic

$$\tilde{Z} \equiv \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$$

[1] A rule of thumb declares n large when

$$np \geq 5 \quad \text{and} \quad n(1-p) \geq 5.$$

is approximately normally distributed with mean 0 and variance 1. As a consequence, the results of Subsection 17.3.1 presented in Table (17.33) apply approximately. However, to use the results concerning the probability β of a type II error and the sample size n presented in Table (17.34) and Formulas (17.35)-(17.37), respectively, we have to use a plausible value of σ or assume a normal distribution for X and resort to the techniques described in Subsection 17.3.2.

17.3.4 Any Sample from a Bernoulli Population with Unknown Success Parameter

Let X be a real random variable describing a binary population, for instance the presidential job approval by U.S. voters. Agreeing that an approval is a success, we can assume that X is Bernoulli distributed with unknown success parameter p. Test concerning p are based on a simple sample $X_1, \ldots, X_n$ of size n drawn from X. For such a sample, we know that the sample sum

$$Z_n \equiv \sum_{k=1}^{n} X_k$$

is binomially distributed with parameters n and p. In addition,

$$\mathbf{E}[Z_n] = np \quad \text{and} \quad \mathbf{D}^2[Z_n] = np(1-p).$$

Consider as the null hypothesis $H_0 : p = p_0$ and as the alternative hypothesis $H_1 : p < p_0$, for a some $p_0 \in (0,1)$ which might represent a minimum consensus threshod. Under this hypothesis it turns out that $Z_n \sim Bin(n, p_0)$, so that $\mathbf{E}[Z_n] = np_0$. Hence, to reject the null hypothesis in the direction of the alternative we need to observe a value of the test statistic significantly smaller than np_0. More specifically, in terms of the probability α of committing a type I error, we can Then, write

$$\alpha \equiv \mathbf{P}(\text{reject } H_0 \mid H_0 \text{ is true}) = \mathbf{P}(Z_n \leq z_{n,\alpha} \mid p = p_0) = \mathbf{P}(Z_n \leq z_{n,\alpha} \mid Z_n \sim Bin(n, p_0)).$$

where $z_{n,\alpha}$ is a suitable critical value. On the other hand, we have

$$\mathbf{P}(Z_n \leq z_{n,\alpha} \mid Bin(n, p_0)) = \sum_{k=0}^{[z_{n,\alpha}]} \binom{n}{k} p_0^k (1-p_0)^{n-k}.$$

Therefore, the probability of committing a type I error α and the critical value $z_{n,\alpha}$ are related by a discrete map. Given $\alpha \in (0,1)$ is not always possible to find areal number $z_{n,\alpha}$ such that

$$\sum_{k=0}^{[z_{n,\alpha}]} \binom{n}{k} p_0^k (1-p_0)^{n-k} = \alpha. \tag{17.39}$$

For instance, there is no $z_{n,\alpha} \in \mathbb{R}$ which can make $\alpha < (1-p_0)^n$. Yet, when $\alpha \geq (1-p_0)^n$ there might be either no solutions or infinite solutions of Equation (17.39). However, for any

Another, more conservative, rule of thumb declares n large when

$$np \geq 10 \quad \text{and} \quad n(1-p) \geq 10.$$

$\alpha \geq (1-p_0)^n$ we can always determine the smallest integer $z^-_{n,\alpha}$ such that

$$\sum_{k=0}^{z^-_{n,\alpha}} \binom{n}{k} p_0^k (1-p_0)^{n-k} \geq \alpha \quad \text{and} \quad \sum_{k=z^-_{n,\alpha}+1}^{n} \binom{n}{k} p_0^k (1-p_0)^{n-k} \geq 1-\alpha. \tag{17.40}$$

Thus, we will reject H_0 in the direction compatible with the alternative hypothesis with a significance level $\blacksquare \geq (1-p_0)^n$ if we observe a value of the test statistic not larger than $z_{n,\alpha}$ solution of Equation (**??**). This corresponds to the rejection region $\left(0, z^-_{n,\alpha}\right)$, where $z^-_{n,\alpha}$ is the smallest integer which solves (**??**). Note that the larger n is the smaller the probability of committing a type I error can be made.

To compute the probability β of committing a type II error corresponding to a probability α of committing a type I error, we need to consider

$$\beta \equiv \mathbf{P}\left(\text{not reject } H_0 \mid H_0 \text{ is false}\right) = \mathbf{P}\left(Z_n > z_{n,\alpha} \mid H_0 \text{ is false}\right),$$

where $z_{n,\alpha}$ is the largest integer which solves (**??**). On the other hand, assuming H_0 false, which means $p \neq p_0$, has different implications according to the alternative hypothesis under consideration. In particular, assuming H_0 false in a direction compatible with the alternative hypothesis $H_1 : p < p_0$ means that true value of the parameter p might be any $p_1 < p_0$. In turn, this means we have to assume the hypothesis

$$H_0^{(1)} : p = p_1$$

true for a mispecified $p_1 < p_0$. Eventually, we will determine β as a function of p_1 given that $Z_n \sim Bin\left(n, p_1\right)$, which is clearly equivalent to assume $H_0^{(1)}$ true. That is

$$\beta\left(p_1\right) = \mathbf{P}\left(Z_n > z^-_{n,\alpha} \mid H_0^{(1)} \text{ is true}\right) = \mathbf{P}\left(Z_n > z^-_{n,\alpha} \mid Z_n \sim Bin\left(n, p_1\right)\right)$$

for any $p_1 < p_0$. We Then, obtain

$$\beta\left(p_1\right) = \sum_{k=z^-_{n,\alpha}+1}^{n} \binom{n}{k} p_1^k \left(1-p_1\right)^{n-k}, \tag{17.41}$$

for any $p_1 < p_0$.

The determination of the p-value requires a simple direct computation. Assuming that the observed value of the statistic Z_n is the integer $z_{n,\alpha}$, the p-value is given by

$$\mathbf{P}\left(Z_n \leq z_{n,\alpha} \mid p = p_0\right) = \mathbf{P}\left(Z_n \leq z_{n,\alpha} \mid Z_n \sim Bin\left(n, p_0\right)\right) = \sum_{k=0}^{z_{n,\alpha}} \binom{n}{k} p_0^k \left(1-p_0\right)^{n-k}.$$

In a similar way we can tackle the cases in which the alternative hypothesis are $H_1 : p > p_0$ or $H_1 : p \neq p_0$. However, with reference to the alternative hypothesis $H_1 : p \neq p_0$ the determination of the rejection region leads to determine the integer values $z^+_{n,a/2}$ and $z^+_{n,a/2}$ of the test statistic such that

$$\sum_{k=0}^{z^-_{n,\alpha/2}} \binom{n}{k} p_0^k (1-p_0)^{n-k} \geq \alpha/2 \quad \text{and} \quad \sum_{k=z^-_{n,\alpha}+1}^{n} \binom{n}{k} p_0^k (1-p_0)^{n-k} \geq 1-\alpha/2 \tag{17.42}$$

and

$$\sum_{k=1}^{z^+_{n,\alpha/2}} \binom{n}{k} p_0^k (1-p_0)^{n-k} \geq 1-\alpha/2, \qquad \sum_{k=z^+_{n,\alpha/2}+1}^{n} \binom{n}{k} p_0^k (1-p_0)^{n-k} \geq \frac{\alpha}{2}. \tag{17.43}$$

Furthermore, the probability β of committing a type II error corresponding to a probability α of committing a type I error is given by

$$\beta(p_1) = \sum_{k=z^-_{n,\alpha/2}+1}^{z^+_{n,\alpha/2}-1} \binom{n}{k} p_1^k (1-p_1)^{n-k} \tag{17.44}$$

for any $p_1 \neq p_0$. In the end, given that observed value of the statistic Z_n is the integer $z_n^- < np_0$ [resp. $z_n^+ > np_0$], in terms of p-value we have to compute

$$\sum_{k=0}^{z_n^-} \binom{n}{k} p_0^k (1-p_0)^{n-k} + \sum_{k=0}^{z_n^+} \binom{n}{k} p_0^k (1-p_0)^{n-k},$$

where z_n^+ [resp. z_n^-] is the integer which allows the best approximation of the equation

$$z_n^+ - np_0 = np_0 - z_n^-. \tag{17.45}$$

17.3.5 Large Sample from a Bernoulli Population with Unknown Success Parameter

The procedure described in Subsection 17.3.4 applies to simple sample of any size n. However, when the size n of the sample is large we can resort on an easier computational procedure, which exploits the Gaussian approximation of a binomial random variable. In fact, assuming that n is large[2] the sample sum

$$Z_n \equiv \sum_{k=1}^{n} X_k$$

is approximately Gaussian distributed with mean $\mu_{Z_n} = np$ and variance $\sigma^2_{Z_n} = np(1-p)$. In turn, the sample mean

$$\bar{X}_n \equiv \frac{1}{n} Z_n$$

is approximately Gaussian distributed with mean $\mu_{\bar{X}_n} = p$ and variance $\sigma^2_{\bar{X}_n} = p(1-p)/n$. In the end, the statistic

$$Z \equiv \frac{\bar{X}_n - p}{\sqrt{p(1-p)/n}}$$

[2] A rule of thumb declares n large when

$$np \geq 5 \quad \text{and} \quad n(1-p) \geq 5.$$

Another, more conservative, rule of thumb declares n large when

$$np \geq 10 \quad \text{and} \quad n(1-p) \geq 10.$$

is approximately Gaussian distributed with mean $\mu_Z = 0$ and variance $\sigma_Z^2 = 1$. In particular, the null hypothesis $H_0 : p = p_0$ implies that

$$Z_0 \equiv \frac{\bar{X}_n - p_0}{\sqrt{p_0 (1 - p_0) / n}} \sim N(0, 1).$$

Therefore, in terms of the critical value(s) and the rejection region corresponding to the probability α of committing a type I error in a direction compatible with each of the alternative hypotheses $H_1 : p < p_0$, $H_1 : p > p_0$, and $H_1 : p \neq p_0$ we can apply Table (17.33) of Subsection 17.3.1.

To compute the probability β of committing a type II error corresponding to a probability α of committing a type I error, the usual argument applies: declaring H_0 false has different implications according to the alternative hypothesis under consideration. For instance, declaring H_0 false in a direction compatible with the alternative hypothesis $H_1 : p < p_0$ leads to assume the hypothesis

$$H_0^{(1)} : p = p_1$$

true for a mispecified $p_1 < p_0$. This implies that we will determine β as a function p_1. Moreover, assuming $H_0^{(1)}$ true, for a mispecified $p_1 > p_0$ implies that it is the statistic

$$Z_1 \equiv \frac{\bar{X}_n - p_1}{\sqrt{p_1 (1 - p_1) / n}}$$

to be normally distributed with mean 0 and variance 1. Note that since we have

$$Z_0 = \frac{\sqrt{p_1 (1 - p_1) / n}}{\sqrt{p_0 (1 - p_0) / n}} Z_1 + \frac{p_1 - p_0}{\sqrt{p_0 (1 - p_0) / n}},$$

the statistic Z_0 is still normally distributed, but with mean $\mu_{Z_0} = \frac{p_1 - p_0}{\sqrt{p_0(1-p_0)/n}}$ and variance $\sigma_{Z_0}^2 = \frac{p_0(1-p_0)/n}{p_1(1-p_1)/n}$. On the other hand, we do not reject H_0 when we observe a realization of $Z_0 \geq z_\alpha^-$. In light of the above arguments, we can write

$$\begin{aligned}
\beta(p_1) &= \mathbf{P}\left(\text{not reject } H_0 \mid H_0^{(1)}\right) \\
&= \mathbf{P}\left(Z_0 \geq z_\alpha^- \mid p = p_1\right) \\
&= \mathbf{P}\left(\frac{\sqrt{p_1 (1 - p_1) / n}}{\sqrt{p_0 (1 - p_0) / n}} Z_1 + \frac{p_1 - p_0}{\sqrt{p_0 (1 - p_0) / n}} \geq z_\alpha^- \mid p = p_1\right) \\
&= \mathbf{P}\left(Z_1 \geq \frac{\sqrt{p_0 (1 - p_0) / n}}{\sqrt{p_1 (1 - p_1) / n}} \left(z_\alpha^- - \frac{p_1 - p_0}{\sqrt{p_0 (1 - p_0) / n}}\right) \mid p = p_1\right) \\
&= \mathbf{P}\left(Z_1 \geq \frac{p_0 - p_1 + z_\alpha^- \sqrt{p_0 (1 - p_0) / n}}{\sqrt{p_1 (1 - p_1) / n}}\right) \\
&= 1 - \Phi\left(\frac{p_0 - p_1 + z_\alpha^- \sqrt{p_0 (1 - p_0) / n}}{\sqrt{p_1 (1 - p_1) / n}}\right).
\end{aligned}$$

which yields a close formula to compute the probability of committing a type II error given a significance level α in a direction compatible with the alternative hypothesis $H_1 : p < p_0$.

Analogous arguments apply when we declare H_0 false in a direction compatible with each of the the alternative hypotheses $H_1 : p > p_0$ and $H_1 : p \neq p_0$. As a consequence we obtain the formulas

$$\beta(p_1) = \Phi\left(\frac{p_0 - p_1 + z_\alpha^+ \sqrt{p_0(1-p_0)/n}}{\sqrt{p_1(1-p_1)/n}}\right)$$

and

$$\beta(p_1) = \Phi\left(\frac{p_0 - p_1 + z_{\alpha/2}^+ \sqrt{p_0(1-p_0)/n}}{\sqrt{p_1(1-p_1)/n}}\right) - \Phi\left(\frac{p_0 - p_1 + z_{\alpha/2}^- \sqrt{p_0(1-p_0)/n}}{\sqrt{p_1(1-p_1)/n}}\right),$$

according to whether the alternative hypothesis is $H_1 : p > p_0$ or $H_1 : p \neq p_0$. Using the normal approximation of a binomial random variable, it is also possible to obtain closed formulas for the determination of the sample size n, which allows for a given probability α of committing a type I error the achievement of a probability β of committing a type II error. Replicating the argument of Subsection 17.3.1, set $\beta \equiv \beta(p_1)$ and determine z_β such that

$$\beta = \Phi(-z_\beta) = 1 - \Phi(z_\beta).$$

Then, we can write

$$1 - \Phi(z_\beta) = 1 - \Phi\left(\frac{p_0 - p_1 + z_\alpha^- \sqrt{p_0(1-p_0)/n}}{\sqrt{p_1(1-p_1)/n}}\right),$$

which implies

$$z_\beta = \frac{p_0 - p_1 + z_\alpha^- \sqrt{p_0(1-p_0)/n}}{\sqrt{p_1(1-p_1)/n}}.$$

By a straightforward computation, from the latter it follows

$$n = \left(\frac{z_\beta \sqrt{p_1(1-p_1)} - z_\alpha^- \sqrt{p_0(1-p_0)}}{p_0 - p_1}\right)^2. \tag{17.46}$$

Similar computation show that when we declare H_0 false in a direction compatible with the the alternative hypotheses $H_1 : p > p_0$ Equation (**??**) still holds true, while in a direction compatible with the the alternative hypotheses $H_1 : p \neq p_0$ we obtain

$$n = \left(\frac{z_\beta \sqrt{p_1(1-p_1)} + z_{\alpha/2}^+ \sqrt{p_0(1-p_0)}}{p_0 - p_1}\right). \tag{17.47}$$

17.4 Hypothesis Testing about a Population Variance

17.4.1 Any Sample from a Gaussian Population

Let X be a real random representing a population. We assume that X is Gaussian distributed with unknown variance σ^2. In symbols $X \sim N(\mu, \sigma^2)$. Since we are interested in testing the variance, we make the null hypothesis in the form $H_0 : \sigma^2 = \sigma_0^2$, where σ_0^2, where σ_0^2 is our candidate true value for the variance of X, and the alternative hypothesis in each of the forms

$H_1 : \sigma^2 < \sigma_0^2$, $H_1 : \sigma^2 > \sigma_0^2$, and $H_1 : \sigma^2 \neq \sigma_0^2$. Let $X_1, \ldots, X_n$ be a simple random sample of size n drawn from X and let S_n^2 be the unbiased variance of X. Since X is Gaussian distributed, independently of the size n of the sample, we know that the statistic $(n-1)\, S_n^2/\sigma^2$ is chi-square distributed with $n-1$ degrees of freedom (see Theorem **??**). In symbols

$$\frac{(n-1)\, S_n^2}{\sigma^2} \sim \chi_{n-1}^2.$$

If we assume the null hypothesis true, that is $\sigma^2 = \sigma_0^2$, Then, also the random variable

$$\chi_{n-1}^2(\sigma_0) \equiv \frac{(n-1)\, S_n^2}{\sigma_0^2}$$

is chi-square distributed with $n-1$ degrees of freedom. Therefore, choosing a significance level $\alpha \in (0,1)$ we can determine the critical value(s) of χ_{n-1}^2 corresponding to each of the alternative hypotheses considered. We report the critical value(s) and the rejection region for the cases $\alpha \equiv 0.10,\ 0.05,\ 0.01$, and $n = 4, 5, 6$, in Table 17.48

null hypothesis $H_0 : \sigma^2 = \sigma_0^2$

alt. hypothesis.	signific. lev.	critical value(s) $n = 4, 5, 6$	reject. region	prob. of type I error
$H_1 : \sigma^2 < \sigma_0^2$	$\alpha \equiv 0.10$	$\chi^2_{3,\alpha,-} = \ldots,\ \chi^2_{4,\alpha,-} = \ldots,\ \chi^2_{5,\alpha,-} = \ldots$	$\left(-\infty, \chi^2_{n-1,\alpha,-}\right)$	$\alpha = \mathbf{P}\left(\chi^2_{n-1} \leq \chi^2_{n-1,\alpha,-}\right)$
	$\alpha \equiv 0.05$	$\chi^2_{3,\alpha,-} = \ldots,\ \chi^2_{4,\alpha,-} = \ldots,\ \chi^2_{5,\alpha,-} = \ldots$		
	$\alpha \equiv 0.01$	$\chi^2_{3,\alpha,-} = \ldots,\ \chi^2_{4,\alpha,-} = \ldots,\ \chi^2_{5,\alpha,-} = \ldots$		
$H_1 : \sigma^2 > \sigma_0^2$	$\alpha \equiv 0.10$	$\chi^2_{3,\alpha,+} = \ldots,\ \chi^2_{4,\alpha,+} = \ldots,\ \chi^2_{5,\alpha,+} = \ldots$	$\left(\chi^2_{n-1,\alpha,+}, +\infty\right)$	$\alpha = \mathbf{P}\left(\chi^2_{n-1} \geq \chi^2_{n-1,\alpha,+}\right)$
	$\alpha \equiv 0.05$	$\chi^2_{3,\alpha,+} = \ldots,\ \chi^2_{4,\alpha,+} = \ldots,\ \chi^2_{5,\alpha,+} = \ldots$		
	$\alpha \equiv 0.01$	$\chi^2_{3,\alpha,+} = \ldots,\ \chi^2_{4,\alpha,+} = \ldots,\ \chi^2_{5,\alpha,+} = \ldots$		
$H_1 : \sigma^2 \neq \sigma_0^2$	$\alpha \equiv 0.10$	$\chi^2_{3,\alpha/2,-} = \ldots,\ \chi^2_{4,\alpha/2,-} = \ldots,\ \chi^2_{5,\alpha/2,-} = \ldots$	$\left(-\infty, \chi^2_{n-1,\alpha/2,-}\right)$	$\alpha/2 = \mathbf{P}\left(\chi^2_{n-1} \leq \chi^2_{n-1,\alpha/2,-}\right)$
		$\chi^2_{3,\alpha/2,+} = \ldots,\ \chi^2_{4,\alpha/2,+} = \ldots,\ \chi^2_{5,\alpha/2,+} = \ldots$	$\cup \left(\chi^2_{n-1,\alpha/2,+}, +\infty\right)$	
	$\alpha \equiv 0.05$	$\chi^2_{3,\alpha/2,-} = \ldots,\ \chi^2_{4,\alpha/2,-} = \ldots,\ \chi^2_{5,\alpha/2,-} = \ldots$		$\alpha/2 = \mathbf{P}\left(\chi^2_{n-1} \geq \chi^2_{n-1,\alpha/2,+}\right)$
		$\chi^2_{3,\alpha/2,+} = \ldots,\ \chi^2_{4,\alpha/2,+} = \ldots,\ \chi^2_{5,\alpha/2,+} = \ldots$		
	$\alpha \equiv 0.01$	$\chi^2_{3,\alpha/2,-} = \ldots,\ \chi^2_{4,\alpha/2,-} = \ldots,\ \chi^2_{5,\alpha/2,-} = \ldots$		
		$\chi^2_{3,\alpha/2,+} = \ldots,\ \chi^2_{4,\alpha/2,+} = \ldots,\ \chi^2_{5,\alpha/2,+} = \ldots$		

(17.48)

Proposition 1196 *In terms of p-value, we reject the null hypothesis whether we have either*

$$\mathbf{P}\left(\chi^2_{n-1} \leq \chi^2_{n-1}(\omega)\right) < \alpha/2 \tag{17.49}$$

or

$$\mathbf{P}\left(\chi^2_{n-1} \geq \chi^2_{n-1}(\omega)\right) < \alpha/2 \tag{17.50}$$

or none of the above Equations (17.49) and (17.50) holds true.

17.4.2 Large Sample from a Non-Gaussian Population

Assume X has unknown variance σ^2 and (unknown) finite central moment of the fourth order μ_4. Since we are interested in testing the variance, we make the null hypothesis in the form $H_0 : \sigma^2 = \sigma_0^2$, where σ_0^2 is our candidate true value for the variance of X, and the alternative hypothesis in each of the forms $H_1 : \sigma^2 < \sigma_0^2$, $H_1 : \sigma^2 > \sigma_0^2$, and $H_1 : \sigma^2 \neq \sigma_0^2$. Let $X_1, \ldots, X_n$ be a simple random sample of size n drawn from X and let $\bar{X}_n$ and S_n^2 be the sample mean and the unbiased variance of X, respectively. We have

$$\mathbf{E}\left[S_n^2\right] = \sigma^2$$

and

$$\mathbf{D}^2\left[S_n^2\right] = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right) = \frac{\sigma^4}{n}\left(\hat{\kappa} - \frac{n-3}{n-1}\right),$$

where $\hat{\kappa} \equiv \mu_4/\sigma^4$ is the standardized kurtosis of X (see Proposition **??**). In addition, the statistics

$$\frac{S_n^2 - \sigma^2}{\sqrt{\left(\sum_{k=1}^{n}\left(X_k - \bar{X}_n\right)^4 - \sigma^4\right)/n}}$$

is asymptotically standard Gaussian distributed (see [**?**, Arnold, S.F. (1990), Mathematical Statistics, Prentice Hall, Upper Saddle ... Bain, L.J. (1978),]). If we assume the null hypothesis true, that is $\sigma^2 = \sigma_0^2$, Then, also the random variable

$$\frac{S_n^2 - \sigma_0^2}{\sqrt{\left(\frac{1}{n}\sum_{k=1}^{n}\left(X_k - \bar{X}_n\right)^4 - \sigma_0^4\right)/n}}$$

is asymptotically standard Gaussian distributed. Therefore, choosing a significance level $\alpha \in (0,1)$, we have the standard rejection regions

$$\left(-\infty, z_\alpha^-\right), \quad \left(z_\alpha^+, +\infty\right), \quad \left(-\infty, z_{\alpha/2}^-\right) \cup \left(z_{\alpha/2}^+, +\infty\right).$$

Consider a data sample $(x_k)_{k=1}^n$ and compute

$$z_n \equiv \frac{s_n^2 - \sigma_0^2}{\sqrt{\left(\frac{1}{n}\sum_{k=1}^{n}\left(x_k - \bar{x}_n\right)^4 - \sigma_0^4\right)/n}},$$

where

$$\bar{x}_n \equiv \frac{1}{n}\sum_{k=1}^{n} x_k \quad \text{and} \quad s_n^2 \equiv \frac{1}{n-1}\sum_{k=1}^{n}\left(x_k - \bar{x}_n\right)^2 .$$

We will reject the null hypothesis in favor of the alternative $H_1 : \sigma^2 < \sigma_0^2$ [resp. $H_1 : \sigma^2 > \sigma_0^2$ or $H_1 : \sigma^2 \neq \sigma_0^2$] according to whether $z_n \in (-\infty, z_\alpha^-)$ [resp. $z_n \in (z_\alpha^+, +\infty)$ or $z_n \in \left(-\infty, z_{\alpha/2}^-\right) \cup \left(z_{\alpha/2}^+, +\infty\right)$].

Chapter 18

Univariate Simple Regression Models

18.1 Introduction

Let us consider two real random variables X and Y on a probability space $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$. Assume there exist a function $f : \mathbb{R} \to \mathbb{R}$ and a real random variable U on Ω such that

$$Y = f(X) + U. \tag{18.1}$$

Definition 1197 *Equation (18.1) is called a* simple univariate regression equation*.*

Definition 1198 *In the vast literature on regressions, the variables X, Y, and U are commonly known with different names. The variable X is called the* independent *or* explanatory variable*; it is also called the* predictor *or* regressor *(*variable*). The variable Y is called the* dependent *or* explained variable*; it is also called the* response *or* regressand *(*variable*). The random variable U is called the* error *or* disturbance *or* noise variable*.*

Definition 1199 *Equation (18.1) is also said to be the* regression equation *of Y against X with error U.*

Definition 1200 *We call the function $f : \mathbb{R} \to \mathbb{R}$ in Equation (18.1) the* regression function *of Y against X and we call the graph of f,*

$$\Gamma_f \equiv \left\{(x, y) \in \mathbb{R}^2 : y = f(x),\ x \in \mathbb{R}\right\},$$

the regression curve *of Y against X.*

There is no loss in generality in assuming that

$$\mathbf{E}[U] = 0, \tag{18.2}$$

namely, the error variable has zero mean. This is a natural characteristic of any unsystematic measurement error, independently of its distribution. Furthermore, if (18.2) were not true, we could rewrite Equation (18.1) in the form

$$Y = \tilde{f}(X) + \tilde{U} \tag{18.3}$$

by introducing the function $\tilde{f} : \mathbb{R} \to \mathbb{R}$ given by

$$\tilde{f}(x) \stackrel{\text{def}}{=} f(x) + \mathbf{E}[U]$$

and the random variable

$$\tilde{U} \stackrel{\text{def}}{=} U - \mathbf{E}[U].$$

The latter would have the same type of distribution as U and zero mean. Eventually, Equation (18.3) would clearly be equivalent to (18.1) and $\tilde{U}$ would satisfy Equation (18.2). In light of these considerations, from now on we will assume implicitly that the noise variable U always satisfies Equation (18.2). Therefore, under the additional assumption that U is independent of X, Equation (18.1) implies

$$\mathbf{E}[Y \mid X] = \mathbf{E}[f(X) + U \mid X] = \mathbf{E}[f(X) \mid X] + \mathbf{E}[U \mid X] = f(X) + \mathbf{E}[U] = f(X).$$

Hence, in Equation (18.1), the random variable $f(X)$ turns out to be the conditional expectation $\mathbf{E}[Y \mid X]$. For instance, assuming also that the random variables X and Y are jointly Gaussian distributed, we obtain

$$f(X) = \mathbf{E}[Y] + Corr(X,Y)\frac{\mathbf{D}[Y]}{\mathbf{D}[X]}(X - \mathbf{E}[X]) = \mathbf{E}[Y] + \frac{Cov(X,Y)}{\mathbf{D}^2[X]}(X - \mathbf{E}[X]) \tag{18.4}$$

(see Equation 10.37). This is a closed form representation of the regression function of Y against X when the noise variable U is independent of X and the random variables X and Y are jointly Gaussian. As important consequences, we obtain that the regression function is linear in the regressor X. That is

$$f(x) = \alpha + \beta x, \tag{18.5}$$

for every $x \in \mathbb{R}$, where

$$\beta \equiv \frac{Cov(X,Y)}{\mathbf{D}^2[X]} \quad \text{and} \quad \alpha \equiv \mathbf{E}[Y] - \beta\mathbf{E}[X]. \tag{18.6}$$

Moreover, the noise variable U, which satisfies

$$U = Y - f(X) = Y - \alpha - \beta X,$$

as linear combination of jointly Gaussian variables is necessarily Gaussian distributed (see Definition 837). We will see that the structure of the coefficients of the linear regression of Y against X given by (18.6) remains the same also when the random variables X and Y are not jontly Gaussian and the noise variable U is simply uncorrelated of X. However, weakening the assumption of jointly Gaussianity, the linear regression of Y against X will no longer be the regression of Y against X, meaning that $f(X) \neq \mathbf{E}[Y \mid X]$, but only one of the possible regressions.

Example 1201 *Given any vector $\mu \equiv (\mu_1, \mu_2)^\intercal \in \mathbb{R}^2$, any matrix $A \equiv (a_{j,k})_{j,k=1}^n \in \mathbb{R}^{2\times 2}$, and any pair of independent standard Gaussian variables Z_1 and Z_2, we know that the random vector $(X,Y)^\intercal$ satisfying the equation*

$$\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix} \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}$$

is Gaussian distributed (see Theorem 838), that is to say the random variables X and Y are jointly Gaussian distributed. Therefore, since we have

$$\mathbf{E}[X] = \mu_1, \quad \mathbf{D}^2[X] = a_{1,1}^2 + a_{1,2}^2, \quad \mathbf{E}[Y] = \mu_2, \quad \mathbf{D}^2[Y] = a_{2,1}^2 + a_{2,2}^2,$$

and

$$Cov(X,Y) = a_{1,1}a_{2,1} + a_{1,2}a_{2,2}$$

(see Proposition 839), regressing Y against X, we obtain that the regression function $f : \mathbb{R} \to \mathbb{R}$ is given by

$$f(x) = \mu_2 + \frac{a_{1,1}a_{2,1} + a_{1,2}a_{2,2}}{a_{1,1}^2 + a_{1,2}^2}(x - \mu_1),$$

for every $x \in \mathbb{R}$ (see 18.6), and the noise variable U is Gaussian distributed.

The goal of the regression analysis is to investigate the structure of the regression function $f : \mathbb{R} \to \mathbb{R}$, in light of some realizations of the variables X and Y and additional assumptions on the noise term U. However, in general, determining the structure of the regression function from scratch is a too ambitious goal. Actually, it is usually postulated a specific structure of the regression function depending on some unknown vector of parameters. That is, it is assumed

$$f(x) \equiv f(x;\theta), \tag{18.7}$$

where $f(x;\theta)$ has specific form and $\theta \equiv (\theta_1, \ldots, \theta_M) \in \Theta \subseteq \mathbb{R}^M$, for some $M \geq 1$, is a vector of real parameters. As a consequence, the goal of the regression analysis reduces to the determination of the best estimate, in a sense that will be made precise below, for the true value of the entries of θ.

Definition 1202 *We call the vector $\theta \equiv (\theta_1, \ldots, \theta_M)$ introduced in Equation (18.7) the* regression parameter vector. *We also refer to the entries of θ as the* regression parameters.

Example 1203 (Trivial Regression) *The simplest example of regression function is the function $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ given by*

$$f(x;\theta) \stackrel{def}{=} \theta, \quad \forall x \in \mathbb{R}, \tag{18.8}$$

where $\theta \in \mathbb{R}$ is the regression parameter.

Definition 1204 (Trivial Regression) *We call a regression function of the form (18.8) a* simple trivial regession *of regression parameter θ and we call the regression curve Γ_f the* regression line *of* intercept *θ.*

Remark 1205 (Trivial Regression) *Assuming that the noise variable U is independent of the explanatory variable X and the regression function f is trivial, we have*

$$\theta \equiv \mathbf{E}[Y]. \tag{18.9}$$

Proof. Since the noise variable U is independent of X we have

$$f(X) = \mathbf{E}[Y \mid X].$$

On the other hand, f is supposed to be trivial, which means

$$f(X) = \theta.$$

We Then, have

$$\theta = \mathbf{E}[f(X)] = \mathbf{E}[\mathbf{E}[Y \mid X]] = \mathbf{E}[Y],$$

as claimed. □

Remark 1206 (Trivial Regression) *Assuming that the noise variable U is independent of explanatory variable X and also the explained variable Y is independent of X, Then, the regression function is necessarily trivial and Equation (18.9) holds true.*

Proof. Since the noise variable U is independent of X, we have

$$f(X) = \mathbf{E}[Y \mid X]. \tag{18.10}$$

On the other hand, since also the explained variable Y is independent of X, we have

$$\mathbf{E}[Y \mid X] = \mathbf{E}[Y]. \tag{18.11}$$

Combining (18.10) and (18.11) it follows that the regression function f is trivial and thanks to Remark 1205 we obtain Equation (18.9). □

Example 1207 (Linear Regression) *The simplest, non trivial, example of regression function is the function $f : \mathbb{R} \times \mathbb{R}^2 \to \mathbb{R}$ given by*

$$f(x;\theta) \overset{def}{=} \alpha + \beta x, \quad \forall x \in \mathbb{R}, \tag{18.12}$$

where $\theta \equiv (\alpha, \beta) \in \mathbb{R}^2$ is the regression parameter.

Definition 1208 (Linear Regression) *We call a regression function of the form (18.12) a* simple linear regression *of regression parameters α and β and we call the regression curve Γ_f the* regression line *of* intercept *α and* slope *β.*

Remark 1209 (Linear Regression) *Assuming that the noise variable U is uncorrelated of the explanatory variable X and the regression function f is linear in symbols*

$$Y = \alpha + \beta X + U, \quad U \perp X,$$

we have

$$\beta = Corr(X,Y)\frac{\mathbf{D}[Y]}{\mathbf{D}[X]} \quad \text{and} \quad \alpha = \mathbf{E}[Y] - \beta \mathbf{E}[X]. \tag{18.13}$$

As a consequence, we have

$$Corr(X,Y) = \frac{\beta}{\sqrt{\beta^2 + \frac{\mathbf{D}^2[U]}{\mathbf{D}^2[X]}}} \quad \text{and} \quad \beta^2 = \frac{Corr(X,Y)^2}{1 - Corr(X,Y)^2}\frac{\mathbf{D}^2[U]}{\mathbf{D}^2[X]}. \tag{18.14}$$

Proof. Since the noise variable U is independent of X, we have

$$\begin{aligned} Cov(X,Y) &= Cov(X, \alpha + \beta X + U) \\ &= \alpha Cov(X, 1_\Omega) + \beta Cov(X,X) + Cov(X,U) \\ &= \beta \mathbf{D}^2[X]. \end{aligned} \tag{18.15}$$

On the other hand,

$$Cov(X,Y) = Corr(X,Y)\,\mathbf{D}[X]\,\mathbf{D}[Y]. \tag{18.16}$$

Hence, combining (18.15)-(18.16), we obtain the desired equation for β. The equation for α trivially follows computing $\mathbf{E}[Y]$. In the end, observing that

$$\mathbf{D}[Y] = \sqrt{\mathbf{D}^2[Y]} = \sqrt{\mathbf{D}^2[\alpha + \beta X + U]} = \sqrt{\beta^2 \mathbf{D}^2[X] + \mathbf{D}^2[U]}, \tag{18.17}$$

from the equation for β we have

$$Corr(X, Y) = \beta \frac{\mathbf{D}[X]}{\sqrt{\beta^2 \mathbf{D}^2[X] + \mathbf{D}^2[U]}}.$$

The latter clearly implies (18.14).

Example 1210 *Other examples of commonly used regression functions are:*

1. *the function* $f : \mathbb{R} \times \mathbb{R}^{M+1} \to \mathbb{R}$ *given by*

$$f(x; \theta) \stackrel{def}{=} \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots + \alpha_M x^M, \quad \forall x \in \mathbb{R}, \tag{18.18}$$

   where $\theta \equiv (\alpha_0, \alpha_1, \ldots, \alpha_M) \in \mathbb{R}^{M+1}$*, for* $M \geq 2$*, is the regression parameter;*

2. *the function* $f : \mathbb{R} \times \mathbb{R}^2 \to \mathbb{R}$ *given by*

$$f(x; \theta) \stackrel{def}{=} \alpha \exp(\beta x), \quad \forall x \in \mathbb{R}, \tag{18.19}$$

   where $\theta \equiv (\alpha, \beta) \in \mathbb{R}^2$ *is the regression parameter;*

3. *the function* $f : \mathbb{R} \times \mathbb{R}^2 \to \mathbb{R}$ *given by*

$$f(x; \theta) \stackrel{def}{=} \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}, \quad \forall x \in \mathbb{R}, \tag{18.20}$$

   where $\theta \equiv (\alpha, \beta) \in \mathbb{R}^2$ *is the regression parameter.*

4. *the function* $f : \mathbb{R} \times \mathbb{R}^{3M} \to \mathbb{R}$ *given by*

$$f(x; \theta) \stackrel{def}{=} \sum_{m=1}^{M} \alpha_m \cos(2\pi\omega_m x) + \sum_{m=1}^{M} \beta_m \cos(2\pi\omega_m x), \quad \forall x \in \mathbb{R}, \tag{18.21}$$

   where $\theta \equiv (\alpha_1, \ldots, \alpha_M, \beta_1, \ldots, \beta_M, \omega_1, \ldots, \omega_M) \in \mathbb{R}^{3M}$.

Definition 1211 (Polynomial Regression) *We call a regression function of the form (18.18), in 1 of Example 1210, the* polynomial regression *of order* m *and regression parameters* α_0, α_1, α_2,...,α_m. *In particular, when* $m = 2$ *we speak of* quadratic regression *and we call the regression curve* Γ_f *the* regression parabola *of* intercept α_0 *and* vertex $(-\alpha_1/2\alpha_2, (4\alpha_0\alpha_2 - \alpha_1^2)/4\alpha_2)$.

Definition 1212 (Exponential Regression) *We call a regression function of the form (18.19), in 2 of Example 1210, the* exponential regression *of regression parameters* α *and* β.

Definition 1213 (Logistic Regression) *We call a regression function of the form (18.20), in 3 of Example 1210, the* logistic regression *of regression parameter* α *and* β.

Definition 1214 (Harmonic Regression) *We call a regression function of the form (18.21), in 4 of Example 1210, the* harmonic regression *of order M and regression* amplitude *parameters $\alpha_1, \ldots, \alpha_M$, $\beta_1, \ldots, \beta_M$ and* frequency *parameters $\omega_1, \ldots, \omega_M$.*

In Finance, Economics, and Medical Sciences [resp. Physics and Engineering] the exponential and the logistic [resp. harmonic] regressions play an importat role. Below, we show some simple examples of regression equations.

Example 1215 (Hooke's Law) *Hooke's Law states that the lenght L of a spring with a fixed extreme stressed or compressed by a force F, in the parallel direction to the spring, is proportional to the intensity of the stressing force. In symbols*

$$L = L_0 + KF, \tag{18.22}$$

where L_0 is the length of the spring at rest and K is the stiffness of the spring. An experimental investigation of Hooke's Law requires considering the unavoidable imperfections in the material constituting the spring and the inaccuracy of the measures of the variables L and F. Hence, from an experimental point of view, Equation (18.22) should be more properly rewritten as

$$L = L_0 + KF + U \tag{18.23}$$

where U is a random variable independent of F which summarizes all possible deviations from (18.22). Equation (18.23) is a linear regression equation for the regressand $Y \equiv L$ against the regressor $X \equiv F$ with intercept parameter $\alpha \equiv L_0$, slope parameter $\beta \equiv K$, and noise term U.

Example 1216 (Ohm's Law) *Ohm's Law states that the intensity I of the electric current between the two ends of a homogeneous conductor is directly proportional to the voltage V across the ends of the conductor. The reciprocal of the constant of proportionality is known as the resistance R of the conductor. In symbols,*

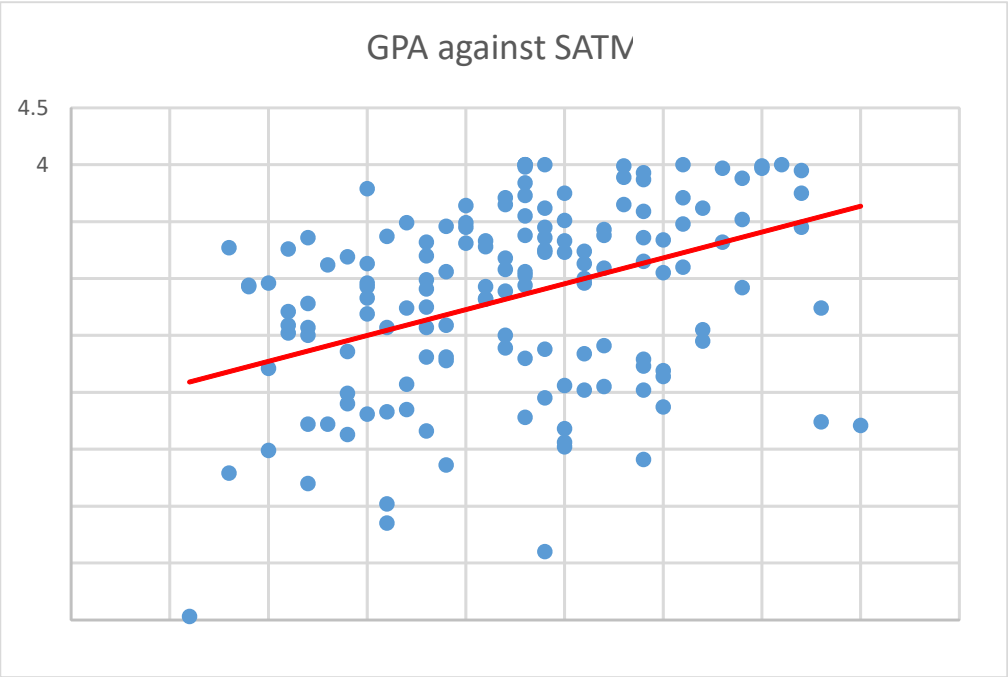
$$I = R^{-1}V. \tag{18.24}$$

As in the case of the Hooke's Law, an experimental investigation of Ohm's Law requires considering the unavoidable imperfections in the material constituting the conductor and the inaccuracy of the measures of the variables I and V. Hence, from an experimental point of view, Equation (18.24) should be more properly rewritten as

$$I = R^{-1}V + U. \tag{18.25}$$

where U is a random variable independent of V which summarizes all possible deviations from (18.24). Equation (18.25) is a linear regression equation for the regressand $Y \equiv I$ against the regressor $X \equiv V$ with intercept parameter $\alpha \equiv 0$, slope parameter $\beta \equiv R^{-1}$, and noise term U.

Example 1217 (GPA against SATM) *Campbell & McCabe (see [2]) considered the problem of predicting the success of freshmen in a Computer Science Mayor on the basis of their performances at the high school level. We report a scatter plot from Campbell & McCabe's data relating the grade pont average (GPA) of a single student in the computer science program, at a large midwestern university, and his/her score on the mathematics section of the scolastic attitude test (SATM). We assume the variable expressing each student's score SATM as the explanatory variable X and the variable expressing the student's GPA as the explained variable Y.*



Differently than Examples (1215) and (1216), in which a linear regression can be easily derived from a well established physical law, in this case no theoretical law is known to relate students' GPA and their score in SATM. Therefore, a possible regression has to established only on the basis of available data. However, looking at Figure 1217 we can recognize a certain linear trend in the distribution of the points in the scatter plot, as evidenced by the trend line plotted. In light of this, we will consider the possibility of a linear regression of Y against X given by

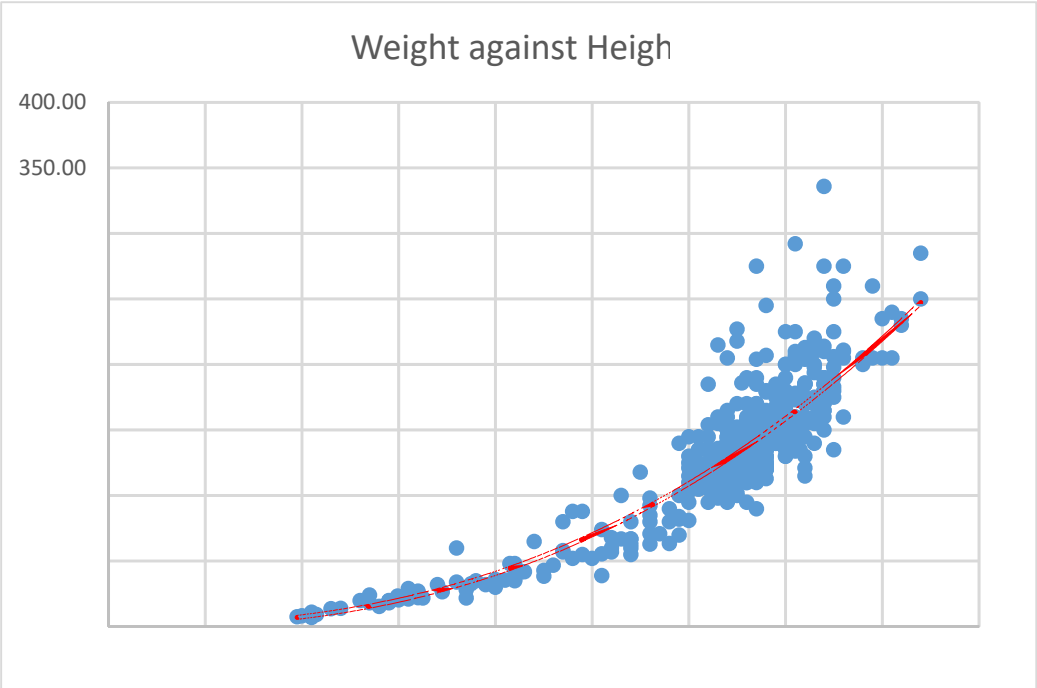
$$Y = \alpha + \beta X + U,$$

where α and β are real parameters and U is the noise term.

Example 1218 (Height against Weight) *Prof. N. Korevaar - University of Utah (see* `https://faculty.utah.edu/u0035213-NICK_KOREVAAR/teaching/index.hml` *) collected a number of human height-weight data which students in the Utah ACCESS class put together from friends and family on summer 2009 (see* `http://www.math.utah.edu/~korevaar/2270fall09/project2/htwts09.pdf`*, see also* `http://www.math.utah.edu/~korevaar/2270fall09/project2/`*). The goal was to relate the height and the weight of a large group of individuals, including infants, to test the body mass index (BMI) hypothesis. The latter claims that the weight of a human being should be roughly proportional to the square of height. Below, we report the scatter plot relating the weight in pounds and the height in inches of the individuals in Korevaar's data base. We consider the variable expressing each individual's height as the explanatory variable X and the variable expressing the individual's weight as the explained variable Y. Looking at Figure 1218 we can recognize a certain quadratic trend in the distribution of the points in the scatter plot, as evidenced by the trend line plotted. In light of this, we will consider the possibility of a quadratic regression of Y against X given by*

$$Y = \beta X^2 + U, \tag{18.26}$$

where β is a real parameter and U is the noise term. Nevertheless, a problem here is that the distance of the points in the scatter plot from the regression parabola appears to increase on the



increasing of the values taken by the random variable X. This makes impausible to validate the assumption that the noise variable U in Equation (18.26) is independent of X.

Example 1219 (Ebola Outbreaks in Guinea) *The following scatter plot is drawn from the data published in WHO Situation Report n=3804 on total suspected, probable, and confirmed cases of Ebola disease during the Ebola outbreak in Guinea in the period March 25, 2014 – February 14, 2016 (see wttps://www.cdc.gov/vhf/ebola/outbreaks/2014-west-africa/cumulative-cases-graphs.html). Here, the explanatory variable X is the day of the recording and the explained variable Y is the number of the recorded cases. Looking at Figure 1219 we can recognize a very strong logistic trend in the distribution of the points in the scatter plot. In light of this, we will consider the possibility of a logistic regression of Y against X given by*

$$Y = \frac{e^{\alpha+\beta X}}{1+e^{\alpha+\beta X}} + U$$

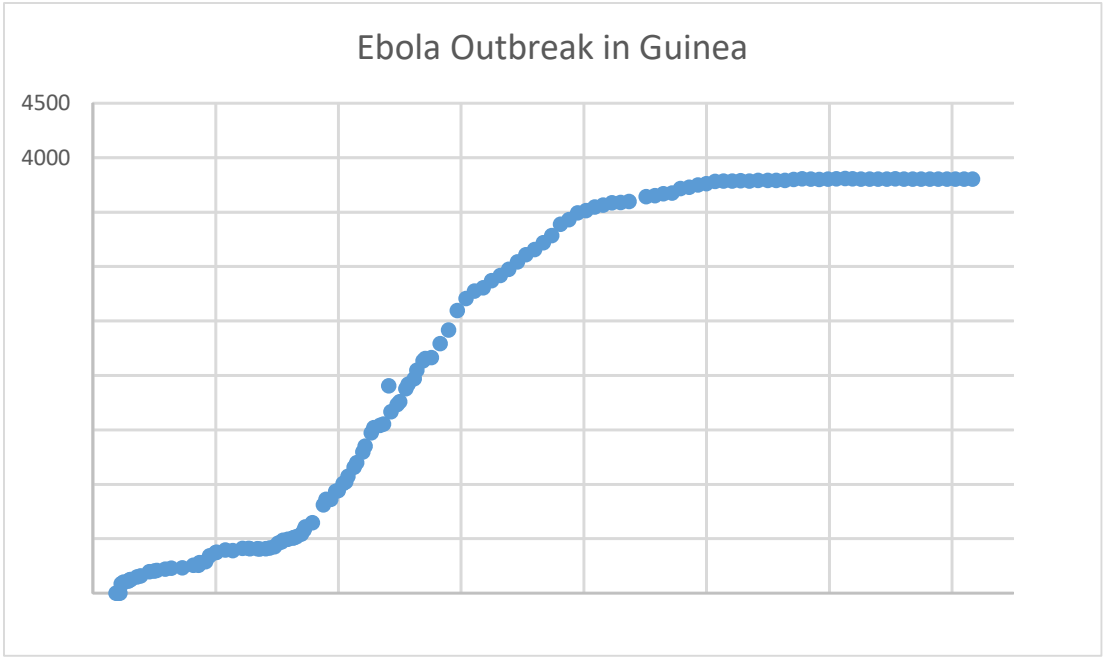
where α and β are real parameters and U is the noise term..

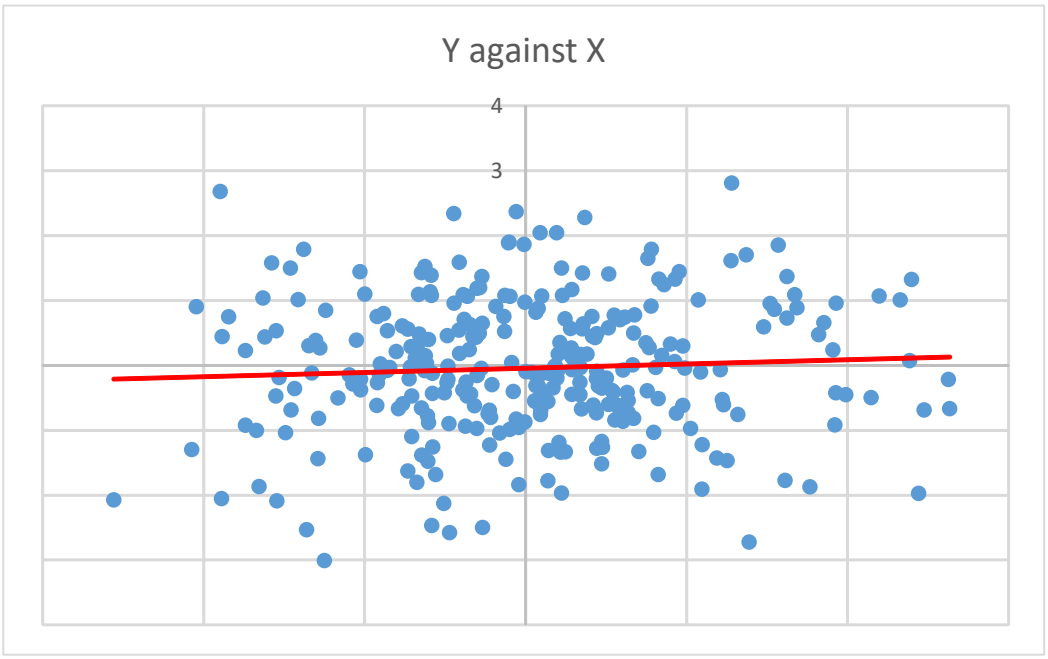
Example 1220 (Independent Gaussian Random Variables) *We have sampled* 300 *values drawn from two independent Gaussian standard random variables X and Y. Below, we report the scatter plot of Y against X. Looking at Figure 1220 we can recognize no trend in the distribution of the points in the scatter plot, but a diffused cloud of points around the origin of the axis with a higher concentration close to the origin. Indeed, the trend line plotted is almost horizontal, which means that no trend is distinguishable. In this case, Equation (18.1) reduces to the trivial equation*

$$Y = \mathbf{E}[Y] + U,$$

as it turns out by applying the conditional expectation operator $\mathbf{E}[\cdot \mid X]$ to both sides of (18.1).

Obviously, in regression analysis it is important to keep out cases like the latter.


Ebola Outbreak in Guinea
4500
4000


Y against X
4
3

18.2 Univariate Regression of Observed Datasets

For some $n \geq 2$ consider n independent measurements of the real variable X [resp. Y] on a probability space $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$, that is a simple random sample $X_1, \ldots, X_n$ [resp. $Y_1, \ldots, Y_n$] of size n drawn from X [resp. Y], and write x_k [resp. y_k] for the value of the variable X [resp. Y] observed at the kth measurement, namely x_k [resp. y_k] is the value $X_k(\omega)$ [resp. $Y_k(\omega)$], for $k = 1, \ldots, n$, on the occureence of some $\omega \in \Omega$.

Definition 1221 *We call the sequence $(x_k)_{k=1}^n$ [resp. $(y_k)_{k=1}^n$] a (simple) dataset of size n drawn from X [resp. Y].*

Assume there exist a function $f : \mathbb{R} \to \mathbb{R}$ and a random variable U independent of X such that Equation (18.1) holds true. In terms of the datasets $(x_k)_{k=1}^n$ and $(y_k)_{k=1}^n$ we Then, derive the equations

$$y_k = f(x_k) + u_k, \quad k = 1, \ldots, n, \tag{18.27}$$

where the sequence $(u_k)_{k=1}^n$ is made by independent realizations of the error random variable U at the kth measurement, for $k = 1, \ldots, n$, namely u_k is the values $U_k(\omega)$, taken by the kth component of a simple random sample $U_1, \ldots, U_n$ drawn from U, on the occurrence of ω.

Definition 1222 *We call Equations (18.27) the observed regression equations of Y against X.*

We stress that in regression analysis we always assume that the dataset $(x_k)_{k=1}^n$ is observed. We will assume that the dataset $(y_k)_{k=1}^n$ is observed or observable, but not yet observed, according to whether we are interested in estimating the parameters of the regression function or studying the properties of some estimators of these parameters. The true values of the parameters of the regression function are to be considered not observed neither observable, as well as the sequence $(u_k)_{k=1}^n$ of independent realizations of the error random variable U. From Equations (18.27), it is clearly seen that despite the assumption that the values of the dataset $(x_k)_{k=1}^n$ are always observed the ignorance of the true values of the parameters of the regression function makes the values of the dataset $(f(x_k))_{k=1}^n$ not observable. Therefore, the observability of the values of the dataset $(y_k)_{k=1}^n$ is not sufficient to observe the sequence of errors $(u_k)_{k=1}^n$. Conversely, the impossibility of observing the the sequence of errors $(u_k)_{k=1}^n$ makes the values of the dataset $(f(x_k))_{k=1}^n$ not observable and prevents the observation of the true values of the parameters of the regression function.

Example 1223 (Hooke's Law) *An investigator considers a homogeneous spring with a fixed extreme. He applies several straining and compressing forces of intensities $x_1, \ldots, x_n$ at the free end of the spring, in the parallel direction to the spring, and measures the corresponding lenght $y_1, \ldots, y_n$ of the spring. The investigator knows that the relationship between the observed datasets $(x_k)_{k=1}^n$ and $(y_k)_{k=1}^n$ is given by*

$$y_k = L_0 + K x_k + u_k, \quad k = 1, \ldots, n. \tag{18.28}$$

where $u_1, \ldots, u_n$ are the unavoidable measurement errors (see Equations (18.23 and (18.27)). However, since the true values of the parameters L_0 and K are are not observable, the investigator cannot use the observed datasets $(x_k)_{k=1}^n$ and $(y_k)_{k=1}^n$ to determine the true values of the errors $u_1, \ldots, u_n$. Conversely, since the true values of the errors $u_1, \ldots, u_n$ cannot be observed, the investigator cannot use the observed datasets $(x_k)_{k=1}^n$ and $(y_k)_{k=1}^n$ to determine the true values of the parameters L_0 and K. He has to estimate them.

Example 1224 (Ohm's law) *An investigator considers a homogeneous conductor with free ends. He applies several voltages $x_1, \ldots, x_n$ across the ends of the conductor, and measures the corresponding intensities $y_1, \ldots, y_n$ of the electric current between the two ends of the conductor. The investigator knows that the relationship between the observed datasets $(x_k)_{k=1}^n$ and $(y_k)_{k=1}^n$ is given by*

$$y_k = R^{-1} x_k + u_k, \quad k = 1, \ldots, n. \tag{18.29}$$

where $u_1, \ldots, u_n$ are the unavoidable measurement errors (see Equations (18.23 and (18.27)). However, as in Example (1223), since the true value of the parameters R is not observable, the investigator cannot use the observed datasets $(x_k)_{k=1}^n$ and $(y_k)_{k=1}^n$ to determine the true values of the errors $u_1, \ldots, u_n$. Conversely, since the true values of the errors $u_1, \ldots, u_n$ cannot be observed, the investigator cannot use the observed datasets $(x_k)_{k=1}^n$ and $(y_k)_{k=1}^n$ to determine the true value of the parameter R. He has to estimate it.

Given that Equations (18.27) hold true and we assumed a specific form $f(x;\theta)$ for the regression function depending on an unknown vector parameter $\theta \in \Theta \subseteq \mathbb{R}^m$, we can introduce the following *ordinary least squares estimate* of the unknown regression parameter vector θ.

Definition 1225 *We call the* sum of squared errors*, acronym* SSE*, the function $SSE : \Theta \to \mathbb{R}_+$ given by*

$$SSE(\theta) \overset{def}{=} \sum_{k=1}^n (y_k - f(x_k;\theta))^2 . \tag{18.30}$$

Be aware that SSE *is also known as the* residual sum of squares *(*RSS*).*

Remark 1226 *We have*

$$SSE(\theta) = \sum_{k=1}^n u_k^2 \tag{18.31}$$

This makes clear the denomination sum of squared errors*.*

Definition 1227 *We call the* ordinary least squares *(OLS)* estimate *of the unknown regression parameter vector θ the vector $\hat{\theta}$ such that*

$$\hat{\theta} = \underset{\theta \in \Theta}{\arg\min} SSE(\theta) . \tag{18.32}$$

The structure of the function $SSE : \Theta \to \mathbb{R}_+$ leads us to consider two different classes of regression models:

- the models characterized by a regression function which may be nonlinear in the regressor but is linear in the unknown parameters or can be made linear in the parameters by a suitable transformation;
- the models characterized by a regression function which is not linear in the unknown parameters and cannot be made linear in the parameters by any transformation without overparametrizing the model.

The first class of models, the so called *intrinsecally linear models*, allow in principle an analytic treatment of the minimization problem (18.32), because the linearity of the regression function in the parameters results in linear first order conditions for the minimization problem. The second class of models lead to nonlinear first order conditions for the minimization problem (18.32), which are often difficult to solve. For the latter models, numerical methods are often applied.

18.3 Univariate Linear Regression of Observed Datasets

For some $n \geq 2$ let $(x_k)_{k=1}^n$ [resp. $(y_k)_{k=1}^n$] be a dataset of size n drawn from the real variable X [resp. Y]. Consider the linear regression function $f : \mathbb{R} \to \mathbb{R}$ given by Equation (18.12) in Example 1207 and assume that Equation (18.1) holds true. Then, Equations (18.27) take the form

$$y_k = \alpha + \beta x_k + u_k, \quad k = 1, \ldots, n, \tag{18.33}$$

where $(\alpha, \beta) \equiv \theta$ is the vector parameter of the linear regression function and $(u_k)_{k=1}^n$ is a sequence of independent measurement errors. As a consequence, the function $SSE : \Theta \to \mathbb{R}_+$ introduced in Definition 1225 takes the forn

$$SSE(\alpha, \beta) = \sum_{k=1}^n (y_k - (\alpha + \beta x_k))^2 \tag{18.34}$$

and the OLS estimate of the unknown regression vector parameter (α, β) becomes the vector $\left(\hat{\alpha}, \hat{\beta}\right)$ such that

$$\left(\hat{\alpha}, \hat{\beta}\right) = \underset{(\alpha,\beta)\in\mathbb{R}^2}{\arg\min} SSE(\alpha, \beta) \tag{18.35}$$

Lemma 1228 *Setting*

$$\bar{x}_n \equiv \frac{1}{n}\sum_{k=1}^n x_k, \quad \bar{y}_n \equiv \frac{1}{n}\sum_{k=1}^n y_k, \tag{18.36}$$

and

$$s_{X,n}^2 \equiv \frac{1}{n-1}\sum_{k=1}^n (x_k - \bar{x}_n)^2, \quad s_{X,Y,n} \equiv \frac{1}{n-1}\sum_{k=1}^n (x_k - \bar{x}_n)(y_k - \bar{y}_n). \tag{18.37}$$

we have

$$s_{X,n}^2 = \frac{1}{n-1}\left(\sum_{k=1}^n x_k^2 - n\bar{x}_n^2\right) \tag{18.38}$$

and

$$s_{X,Y,n} = \frac{1}{n-1}\left(\sum_{k=1}^n x_k y_k - n\bar{x}_n\bar{y}_n\right). \tag{18.39}$$

Proof. A straightforward computation yields

$$\begin{aligned}
\sum_{k=1}^n (x_k - \bar{x}_n)^2 &= \sum_{k=1}^n \left(x_k - \frac{1}{n}\sum_{\ell=1}^n x_\ell\right)^2 \\
&= \sum_{k=1}^n \left(x_k^2 - \frac{2}{n}x_k\sum_{\ell=1}^n x_\ell + \frac{1}{n^2}\left(\sum_{\ell=1}^n x_\ell\right)^2\right) \\
&= \sum_{k=1}^n x_k^2 - \frac{2}{n}\sum_{k=1}^n\sum_{\ell=1}^n x_k x_\ell + \frac{1}{n^2}\sum_{k=1}^n\left(\sum_{\ell=1}^n x_\ell\right)^2 \\
&= \sum_{k=1}^n x_k^2 - \frac{2}{n}\left(\sum_{k=1}^n x_k\right)^2 + \frac{1}{n}\left(\sum_{\ell=1}^n x_\ell\right)^2 \\
&= \sum_{k=1}^n x_k^2 - n\left(\frac{1}{n}\sum_{k=1}^n x_k\right)^2 \\
&= \sum_{k=1}^n x_k^2 - n\bar{x}_n^2
\end{aligned}$$

and

$$\begin{aligned}
&\sum_{k=1}^{n}\left(x_k-\bar{x}_n\right)\left(y_k-\bar{y}_n\right)\\
&=\sum_{k=1}^{n}\left(x_k-\frac{1}{n}\textstyle\sum_{\ell=1}^{n}x_\ell\right)\left(y_k-\frac{1}{n}\textstyle\sum_{\ell=1}^{n}y_\ell\right)\\
&=\sum_{k=1}^{n}\left(x_ky_k-\frac{1}{n}x_k\textstyle\sum_{\ell=1}^{n}y_\ell-\frac{1}{n}y_k\textstyle\sum_{\ell=1}^{n}x_\ell+\frac{1}{n^2}\left(\textstyle\sum_{\ell=1}^{n}x_\ell\right)\left(\textstyle\sum_{\ell=1}^{n}y_\ell\right)\right)\\
&=\textstyle\sum_{k=1}^{n}x_ky_k-\frac{1}{n}\sum_{k=1}^{n}\left(x_k\sum_{\ell=1}^{n}y_\ell\right)-\frac{1}{n}\sum_{k=1}^{n}\left(y_k\sum_{\ell=1}^{n}x_\ell\right)+\frac{1}{n^2}\sum_{k=1}^{n}\left(\sum_{\ell=1}^{n}x_\ell\right)\left(\sum_{\ell=1}^{n}y_\ell\right)\\
&=\textstyle\sum_{k=1}^{n}x_ky_k-\frac{2}{n}\left(\sum_{k=1}^{n}x_k\right)\left(\sum_{\ell=1}^{n}y_\ell\right)+\frac{1}{n}\left(\sum_{\ell=1}^{n}x_\ell\right)\left(\sum_{\ell=1}^{n}y_\ell\right)\\
&=\textstyle\sum_{k=1}^{n}x_ky_k-n\left(\frac{1}{n}\sum_{k=1}^{n}x_k\right)\left(\frac{1}{n}\sum_{k=1}^{n}y_k\right)\\
&=\textstyle\sum_{k=1}^{n}x_ky_k-n\bar{x}_n\bar{y}_n.
\end{aligned}$$

The desired (18.38) and (18.39) immediately follow. □

Corollary 1229 *We have*

$$\frac{s_{X,Y,n}}{s_{X,n}^2}=\frac{\sum_{k=1}^{n}x_ky_k-n\bar{x}_n\bar{y}_n}{\sum_{k=1}^{n}x_k^2-n\bar{x}_n^2}. \tag{18.40}$$

Proof. Combining (18.38) and (18.39), we clearly obtain Equation (18.40). □

Proposition 1230 *With reference to the notation of Lemma 1228, the OLS estimates $\hat{\alpha}$ and $\hat{\beta}$ of the regression parameters α and β are given by*

$$\hat{\alpha}\equiv\bar{y}_n-\hat{\beta}\;\bar{x}_n\quad\textit{and}\quad\hat{\beta}\equiv\frac{s_{X,Y,n}}{s_{X,n}^2}. \tag{18.41}$$

Proof. The OLS estimates in (18.41) are to be obtained by the minimization of the function $SSE:\mathbb{R}^2\to\mathbb{R}_+$ given by (1225). To this, we consider the first order conditions

$$\partial_\alpha SSE\left(\alpha,\beta\right)=\partial_\beta SSE\left(\alpha,\beta\right)=0.$$

Rewriting

$$\begin{aligned}
SSE\left(\alpha,\beta\right)&=\textstyle\sum_{k=1}^{n}y_k^2+n\alpha^2+\beta^2\sum_{k=1}^{n}x_k^2-2\alpha\sum_{k=1}^{n}y_k-2\beta\sum_{k=1}^{n}x_ky_k+2\alpha\beta\sum_{k=1}^{n}x_k\\
&=\textstyle\sum_{k=1}^{n}y_k^2+n\alpha^2+\beta^2\sum_{k=1}^{n}x_k^2-2\alpha n\bar{y}_n-2\beta\sum_{k=1}^{n}x_ky_k+2\alpha\beta n\bar{x}_n,
\end{aligned}$$

we obtain

$$\partial_\alpha SSE\left(\alpha,\beta\right)=2n\alpha-2n\bar{y}_n+2n\beta\bar{x}_n$$

and

$$\partial_\beta SSE\left(\alpha,\beta\right)=2\beta\textstyle\sum_{k=1}^{n}x_k^2-2\sum_{k=1}^{n}x_ky_k+2\alpha n\bar{x}_n.$$

Therefore, the first order conditions yield

$$\alpha+\bar{x}_n\beta=\bar{y}_n \tag{18.42}$$

and

$$n\bar{x}_n\alpha + \left(\sum_{k=1}^n x_k^2\right)\beta = \sum_{k=1}^n x_k y_k, \tag{18.43}$$

respectively. From (18.42), we have

$$\alpha = \bar{y}_n - \bar{x}_n\beta \tag{18.44}$$

and replacing (18.44) into (18.43), it follows

$$\left(\sum_{k=1}^n x_k^2 - n\bar{x}_n^2\right)\beta = \sum_{k=1}^n x_k y_k - n\bar{x}_n\bar{y}_n. \tag{18.45}$$

Equation (18.45) implies

$$\beta = \frac{\sum_{k=1}^n x_k y_k - n\bar{x}_n\bar{y}_n}{\sum_{k=1}^n x_k^2 - n\bar{x}_n^2}. \tag{18.46}$$

Hence, combining (18.44) and (18.46), on account of (18.40), we obtain the desired (18.41) as a candidate minimum for SSE. In the end, observe that we have

$$\partial^2_{\alpha,\alpha}SSE(\alpha,\beta) = 2n, \quad \partial^2_{\alpha,\beta}SSE(\alpha,\beta) = 2n\bar{x}_n, \quad \partial^2_{\beta,\beta}SSE(\alpha,\beta) = 2\sum_{k=1}^n x_k^2.$$

It follows that the Hessian $HSSE$ of SSE is given by

$$HSSE(\alpha,\beta) \equiv \begin{pmatrix} \partial^2_{\alpha,\alpha}SSE(\alpha,\beta) & \partial^2_{\alpha,\beta}SSE(\alpha,\beta) \\ \partial^2_{\alpha,\beta}SSE(\alpha,\beta) & \partial^2_{\beta,\beta}SSE(\alpha,\beta) \end{pmatrix} = 2\begin{pmatrix} n & n\bar{x}_n \\ n\bar{x}_n & \sum_{k=1}^n x_k^2 \end{pmatrix},$$

for every $(\alpha,\beta) \in \mathbb{R}^2$. Therefore, on account of (18.38), we have

$$\det\left(\frac{1}{2}HSSE\right) = n\left(\sum_{k=1}^n x_k^2 - n\bar{x}_n^2\right) = n(n-1)s_{X.n}^2$$

and

$$\operatorname{trace}\left(\frac{1}{2}HSSE\right) = n + \sum_{k=1}^n x_k^2.$$

As a consequence, for $n \geq 2$, the eigenvalues of Hg are always strictly positive, no matter of the choice of (α,β). This implies that the function SSE is convex and the candidate minimum is actually the minimum of SSE. □

Definition 1231 *We call the line in the cartesian plane $\mathbb{R}^2$ of equation*

$$y = \hat{\alpha} + \hat{\beta}x$$

the regression line *of the data sets $(x_k)_{k=1}^n$ and $(y_k)_{k=1}^n$. We call the parameter $\hat{\alpha}$ [resp. $\hat{\beta}$] given by (18.41) the* intercept *[resp.* slope*] OLS* sample estimate *of the regression line.*

Proposition 1232 *Consider the sequence $(u_k)_{k=1}^n$ of independent measurement errors in Equation (18.33). Then, the relationship between the OLS estimates $\hat{\alpha}$ and $\hat{\beta}$ of the intercept and slope parameters of the regression line and the true values α and β are given by*

$$\hat{\alpha} = \alpha + \frac{\frac{1}{n}\sum_{k=1}^n u_k \sum_{k=1}^n x_k^2 - \bar{x}_n \sum_{k=1}^n x_k u_k}{\sum_{k=1}^n x_k^2 - n\bar{x}_n^2} \tag{18.47}$$

and

$$\hat{\beta} = \beta + \frac{\sum_{k=1}^n x_k u_k - \bar{x}_n \sum_{k=1}^n u_k}{\sum_{k=1}^n x_k^2 - n\bar{x}_n^2}. \tag{18.48}$$

Proof. On account of (18.40) and (18.41), we can write

$$\begin{aligned}\hat{\alpha} &= \bar{y}_n - \frac{s_n(x,y)}{s_{X,n}^2}\bar{x}_n = \bar{y}_n - \frac{\sum_{k=1}^n x_k y_k - n\bar{x}_n\bar{y}_n}{\sum_{k=1}^n x_k^2 - n\bar{x}_n^2}\bar{x}_n \\ &= \frac{\bar{y}_n\left(\sum_{k=1}^n x_k^2 - n\bar{x}_n^2\right) - \left(\sum_{k=1}^n x_k y_k - n\bar{x}_n\bar{y}_n\right)\bar{x}_n}{\sum_{k=1}^n x_k^2 - n\bar{x}_n^2}. \end{aligned} \tag{18.49}$$

On the other hand, a straightforward computation yields

$$\begin{aligned}&\bar{y}_n\left(\textstyle\sum_{k=1}^n x_k^2 - n\bar{x}_n^2\right) - \left(\sum_{k=1}^n x_k y_k - n\bar{x}_n\bar{y}_n\right)\bar{x}_n \\ &= \bar{y}_n\textstyle\sum_{k=1}^n x_k^2 - n\bar{x}_n^2\bar{y}_n - \bar{x}_n\sum_{k=1}^n x_k y_k + n\bar{x}_n^2\bar{y}_n \\ &= \frac{1}{n}\textstyle\sum_{\ell=1}^n y_\ell\sum_{k=1}^n x_k^2 - \bar{x}_n\sum_{k=1}^n x_k y_k \\ &= \frac{1}{n}\textstyle\sum_{\ell=1}^n(\alpha+\beta x_\ell + u_\ell)\sum_{k=1}^n x_k^2 - \bar{x}_n\sum_{k=1}^n x_k(\alpha+\beta x_k+u_k) \\ &= \frac{1}{n}\alpha\textstyle\sum_{\ell=1}^n\sum_{k=1}^n x_k^2 + \frac{1}{n}\beta\sum_{\ell=1}^n x_\ell\sum_{k=1}^n x_k^2 + \frac{1}{n}\sum_{\ell=1}^n u_\ell\sum_{k=1}^n x_k^2 \\ &- \left(\alpha\bar{x}_n\textstyle\sum_{k=1}^n x_k + \beta\bar{x}_n\sum_{k=1}^n x_k^2 + \bar{x}_n\sum_{k=1}^n x_k u_k\right) \\ &= \alpha\textstyle\sum_{k=1}^n x_k^2 + \beta\bar{x}_n\sum_{k=1}^n x_k^2 + \frac{1}{n}\sum_{s=1}^n u_s\sum_{k=1}^n x_k^2 - \left(\alpha n\bar{x}_n^2 + \beta\bar{x}_n\sum_{k=1}^n x_k^2 + \bar{x}_n\sum_{k=1}^n x_k u_k\right) \\ &= \alpha\left(\textstyle\sum_{k=1}^n x_k^2 - n\bar{x}_n^2\right) + \frac{1}{n}\sum_{k=1}^n u_k\sum_{k=1}^n x_k^2 - \bar{x}_n\sum_{k=1}^n x_k u_k. \end{aligned} \tag{18.50}$$

Combining (18.49) and (18.50), the desired (18.47) clearly follows. Similarly, we have

$$\hat{\beta} = \frac{s_{X,Y,n}}{s_{X,n}^2} = \frac{\sum_{k=1}^n x_k y_k - n\bar{x}_n\bar{y}_n}{\sum_{k=1}^n x_k^2 - n\bar{x}_n^2}, \tag{18.51}$$

where

$$\begin{aligned}\textstyle\sum_{k=1}^n x_k y_k - n\bar{x}_n\bar{y}_n &= \textstyle\sum_{k=1}^n x_k y_k - \bar{x}_n\sum_{k=1}^n y_k \\ &= \textstyle\sum_{k=1}^n x_k(\alpha+\beta x_k+u_k) - \bar{x}_n\sum_{k=1}^n(\alpha+\beta x_k+u_k) \\ &= \textstyle\sum_{k=1}^n\left(\alpha x_k + \beta x_k^2 + x_k u_k\right) - \bar{x}_n\left(n\alpha + \beta\sum_{k=1}^n x_k + \sum_{k=1}^n u_k\right) \\ &= n\alpha\bar{x}_n + \beta\textstyle\sum_{k=1}^n x_k^2 + \sum_{k=1}^n x_k u_k - n\alpha\bar{x}_n - n\beta\bar{x}_n^2 - \bar{x}_n\sum_{k=1}^n u_k \\ &= \beta\left(\textstyle\sum_{k=1}^n x_k^2 - n\bar{x}_n^2\right) + \sum_{k=1}^n x_k u_k - \bar{x}_n\sum_{k=1}^n u_k. \end{aligned} \tag{18.52}$$

Therefore, combining (18.51) and (18.52), we obtain (18.48). □

Definition 1233 *We call the kth* OLS estimated *or* fitted value *of the dependent variable the real number $\hat{y}_k$ given by*

$$\hat{y}_k \overset{def}{=} \hat{\alpha} + \hat{\beta}x_k, \quad \forall k = 1,\ldots,n. \tag{18.53}$$

Definition 1234 *We call the kth* OLS observed residual *the real number $\hat{u}_k$ given by*

$$\hat{u}_k \overset{def}{=} y_k - \hat{y}_k, \quad \forall k = 1,\ldots,n. \tag{18.54}$$

Note that, both the real numbers $\hat{y}_k$ and $\hat{u}_k$ are observable for any $k = 1, \ldots, n$. From a graphical point of view, the kth estimated value $\hat{y}_k$ of the dependent variable is the ordinate of the point on the regression line with abscissa x_k and the kth residual is the vertical deviation of the point (x_k, y_k) from the regression line. If the residuals are small in magnitude, Then, much of the variability in the data set $(y_k)_{k=1}^n$ can be explained in terms of the variability in the data set $(x_k)_{k=1}^n$ and the linear relationship between the random variables X and Y.

Proposition 1235 *We have*

$$\sum_{k=1}^n \hat{u}_k = 0. \tag{18.55}$$

Proof. Combining Equations (18.53) and (18.54), we can write

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^n \hat{u}_k &= \frac{1}{n}\sum_{k=1}^n (y_k - \hat{y}_k) = \frac{1}{n}\sum_{k=1}^n \left(y_k - \left(\hat{\alpha} + \hat{\beta}\ x_k\right)\right) \\ &= \frac{1}{n}\sum_{k=1}^n y_k - \frac{1}{n}\sum_{k=1}^n \hat{\alpha} - \frac{1}{n}\sum_{k=1}^n \hat{\beta}\ x_k \\ &= \bar{y}_n - \hat{\alpha} - \bar{x}_n\hat{\beta}.\end{aligned}$$

On the other hand, by Equation 18.41, it follows

$$\frac{1}{n}\sum_{k=1}^n \hat{u}_k = 0,$$

which clearly implies the desired result. □

Proposition 1236 *We have*

$$\frac{1}{n}\sum_{k=1}^n \hat{y}_k = \bar{y}_n. \tag{18.56}$$

Proof. We just observe that

$$\frac{1}{n}\sum_{k=1}^n \hat{y}_k - \bar{y}_n = \frac{1}{n}\sum_{k=1}^n \hat{y}_k - \frac{1}{n}\sum_{k=1}^n y_k = \frac{1}{n}\sum_{k=1}^n (\hat{y}_k - y_k) = \frac{1}{n}\sum_{k=1}^n \hat{u}_k = 0.$$

as claimed. □

Proposition 1237 *We have*

$$\sum_{k=1}^n \hat{u}_k x_k = 0. \tag{18.57}$$

Proof. Thanks to (18.40) and (18.41), we can write

$$\begin{aligned}\sum_{k=1}^n \hat{u}_k x_k &= \sum_{k=1}^n (y_k - \hat{y}_k)\, x_k \\ &= \sum_{k=1}^n \left(y_k - \left(\hat{\alpha} + \hat{\beta}x_k\right)\right) x_k \\ &= \sum_{k=1}^n y_k x_k - \hat{\alpha}\sum_{k=1}^n x_k - \hat{\beta}\sum_{k=1}^n x_k^2 \\ &= \sum_{k=1}^n y_k x_k - \hat{\alpha} n\ \bar{x}_n - \hat{\beta}\sum_{k=1}^n x_k^2 \\ &= \sum_{k=1}^n y_k x_k - \left(\bar{y}_n - \hat{\beta}\bar{x}_n\right) n\ \bar{x}_n - \hat{\beta}\sum_{k=1}^n x_k^2 \\ &= \sum_{k=1}^n y_k x_k - n\bar{x}_n\bar{y}_n - \hat{\beta}\left(\sum_{k=1}^n x_k^2 - n\bar{x}_n^2\right) \\ &= 0.\end{aligned}$$

as claimed. □

Proposition 1238 *We have*

$$\hat{u}_k = \frac{\sum_{\ell=1}^{n}\left((n-1)\,s_{X,n}^2\left(\delta_{k,\ell}-\frac{1}{n}\right)-(x_k-\bar{x}_n)\,(x_\ell-\bar{x}_n)\right)y_\ell}{\sum_{k=1}^{n}x_k^2-n\bar{x}_n^2},\quad \forall k=1,\ldots,n,$$

where $\delta_{k,\ell}$ is the Kronecker delta, for all $k,\ell=1,\ldots,n$.

Proof. On account of (18.38), (18.39), (Equation 18.41), and (18.40), we can write

$$\begin{aligned}
\hat{u}_k &= y_k-\hat{y}_k = y_k-\left(\hat{\alpha}+\hat{\beta}x_k\right)\\
&= \left(y_k-\frac{s_{X,Y,n}^2}{s_{X,n}^2}x_k\right)-\left(\bar{y}_n-\frac{s_{X,Y,n}^2}{s_{X,n}^2}\bar{x}_n\right)\\
&= \left(y_k-\frac{\sum_{k=1}^{n}x_ky_k-n\bar{x}_n\bar{y}_n}{(n-1)\,s_{X,n}^2}x_k\right)-\left(\frac{1}{n}\sum_{k=1}^{n}y_k-\frac{\sum_{k=1}^{n}x_ky_k-n\bar{x}_n\bar{y}_n}{(n-1)\,s_{X,n}^2}\bar{x}_n\right)\\
&= \left(y_k-\frac{\sum_{k=1}^{n}x_ky_k-\bar{x}_n\sum_{k=1}^{n}y_k}{(n-1)\,s_{X,n}^2}x_k\right)-\left(\frac{1}{n}\sum_{k=1}^{n}y_k-\frac{\sum_{k=1}^{n}x_ky_k-\bar{x}_n\sum_{k=1}^{n}y_k}{(n-1)\,s_{X,n}^2}\bar{x}_n\right)\\
&= \left(\sum_{\ell=1}^{n}\delta_{k,\ell}y_\ell-\frac{\sum_{\ell=1}^{n}x_\ell y_\ell-\sum_{\ell=1}^{n}\bar{x}_ny_\ell}{(n-1)\,s_{X,n}^2}x_k\right)-\left(\sum_{\ell=1}^{n}\frac{1}{n}y_\ell-\frac{\sum_{\ell=1}^{n}x_\ell y_\ell-\sum_{\ell=1}^{n}\bar{x}_ny_\ell}{(n-1)\,s_{X,n}^2}\bar{x}_n\right)\\
&= \left(\frac{(n-1)\,s_{X,n}^2\sum_{\ell=1}^{n}\delta_{k,\ell}y_\ell-x_k\sum_{\ell=1}^{n}(x_\ell-\bar{x}_n)\,y_\ell}{(n-1)\,s_{X,n}^2}\right)-\left(\frac{(n-1)\,s_{X,n}^2\sum_{\ell=1}^{n}\frac{1}{n}y_\ell-\bar{x}_n\sum_{\ell=1}^{n}(x_\ell-\bar{x}_n)\,y_\ell}{(n-1)\,s_{X,n}^2}\right)\\
&= \frac{\sum_{\ell=1}^{n}(n-1)\,s_{X,n}^2\delta_{k,\ell}y_\ell-\sum_{\ell=1}^{n}x_k\,(x_\ell-\bar{x}_n)\,y_\ell}{\sum_{k=1}^{n}x_k^2-n\bar{x}_n^2}-\frac{\sum_{\ell=1}^{n}\frac{n-1}{n}s_{X,n}^2y_\ell-\sum_{\ell=1}^{n}(x_\ell-\bar{x}_n)\,\bar{x}_ny_\ell}{\sum_{k=1}^{n}x_k^2-n\bar{x}_n^2}\\
&= \frac{\sum_{\ell=1}^{n}\left((n-1)\,s_{X,n}^2\delta_{k,\ell}-x_k\,(x_\ell-\bar{x}_n)\right)y_\ell}{\sum_{k=1}^{n}x_k^2-n\bar{x}_n^2}-\frac{\sum_{\ell=1}^{n}\left(\frac{n-1}{n}s_{X,n}^2-(x_\ell-\bar{x}_n)\,\bar{x}_n\right)y_\ell}{\sum_{k=1}^{n}x_k^2-n\bar{x}_n^2}\\
&= \frac{\sum_{\ell=1}^{n}\left((n-1)\,s_{X,n}^2\delta_{k,s}-x_k\,(x_\ell-\bar{x}_n)-\frac{n-1}{n}s_{X,n}^2+(x_\ell-\bar{x}_n)\,\bar{x}_n\right)y_\ell}{\sum_{k=1}^{n}x_k^2-n\bar{x}_n^2}\\
&= \frac{\sum_{\ell=1}^{n}\left((n-1)\,s_{X,n}^2\left(\delta_{k,s}-\frac{1}{n}\right)-(x_k-\bar{x}_n)\,(x_\ell-\bar{x}_n)\right)y_\ell}{\sum_{k=1}^{n}x_k^2-n\bar{x}_n^2},
\end{aligned}$$

as desired. □

Definition 1239 *We call the* total sum of squares*, acronym* TSS*, or* total variation *in the data set $(y_k)_{k=1}^n$, the sum of the squared deviations of $(y_k)_{k=1}^n$ about the horizontal line of equation $y=\bar{y}_n$, that is the positive number*

$$TSS \overset{def}{=} \sum_{k=1}^{n}(y_k-\bar{y}_n)^2. \tag{18.58}$$

The total sum of squares, TSS, expresses the overall variability in the data set $(y_k)_{k=1}^n$. In fact

Remark 1240 *We have*

$$TSS=(n-1)\,s_{Y,n}^2. \tag{18.59}$$

Proposition 1241 *We have*

$$TSS = \sum_{k=1}^{n} (y_k - \hat{y}_k)^2 + \sum_{k=1}^{n} (\hat{y}_k - \bar{y}_n)^2 . \tag{18.60}$$

Proof. We can write

$$\begin{aligned} TSS &= \sum_{k=1}^{n} (y_k - \hat{y}_k + \hat{y}_k - \bar{y}_n)^2 \\ &= \sum_{k=1}^{n} (y_k - \hat{y}_k)^2 + \sum_{k=1}^{n} (\hat{y}_k - \bar{y}_n)^2 + 2\sum_{k=1}^{n} (y_k - \hat{y}_k)(\hat{y}_k - \bar{y}_n) . \end{aligned} \tag{18.61}$$

On the other hand, we have

$$\hat{y}_k - \bar{y}_n = \hat{\beta}(x_k - \bar{x}_n) \tag{18.62}$$

for every $k = 1, \ldots, n$. In fact, Equation (Equation 18.41) implies

$$\hat{y}_k - \bar{y}_n = \alpha + \hat{\beta}x_k - \bar{y}_n = \bar{y}_n - \hat{\beta}\bar{x}_n + \hat{\beta}x_k - \bar{y}_n = \hat{\beta}(x_k - \bar{x}_n) .$$

Therefore, considering (18.62), (18.55), and (18.57), we obtain

$$\begin{aligned} \sum_{k=1}^{n} (y_k - \hat{y}_k)(\hat{y}_k - \bar{y}_n) &= \hat{\beta}\sum_{k=1}^{n} (y_k - \hat{y}_k)(x_k - \bar{x}_n) \\ &= \hat{\beta}\left(\sum_{k=1}^{n} (y_k - \hat{y}_k) x_k - \bar{x}_n \sum_{k=1}^{n} (y_k - \hat{y}_k)\right) \\ &= \hat{\beta}\left(\sum_{k=1}^{n} \hat{u}_k x_k - \bar{x}_n \sum_{k=1}^{n} \hat{u}_k\right) \\ &= 0. \end{aligned} \tag{18.63}$$

Combining (18.61) and (18.63), Equation (18.60) clearly follows. □

Definition 1242 *We call the* explained sum of squares*, acronym* ESS*, or* explained variation *the positive number*

$$ESS \stackrel{def}{=} \sum_{k=1}^{n} (\hat{y}_k - \bar{y}_n)^2 . \tag{18.64}$$

The explained sum of squares, ESS, can be interpreted as *the amout of the total variation in the data set* $(y_k)_{k=1}^n$ *which can be explained in terms of the variability of the data set* $(x_k)_{k=1}^n$ *through the linear model*. This interpretation is better highlighted by the following Remark

Remark 1243 *We have*

$$ESS = \hat{\beta}^2 (n-1) s_{X,n}^2. \tag{18.65}$$

Proof. In fact, thanks to Equation (18.62), we can write

$$ESS = \sum_{k=1}^{n} \hat{\beta}^2 (x_k - \bar{x}_n)^2 = \hat{\beta}^2 \sum_{k=1}^{n} (x_k - \bar{x}_n)^2 = \hat{\beta}^2 (n-1) s_{X,n}^2,$$

as desired. □

Equation (18.65) with (18.59).

Definition 1244 *We call the* residual sum of squares*, acronym RSS, or* unexplained variation *the sum of the squared deviations about the regression line, that is the positive number*

$$RSS \stackrel{def}{=} \sum_{k=1}^{n} \hat{u}_k^2 \equiv \sum_{k=1}^{n} (y_k - \hat{y}_k)^2 \equiv \sum_{k=1}^{n} \left(y_k - \left(\hat{\alpha} + \hat{\beta}x_k\right)\right)^2 . \tag{18.66}$$

Be aware that RSS is also known as sum of squared residuals *(*SSR*) or* sum of squared estimate of errors *(*SSEE*).*

Proposition 1245 *We have*

$$RSS = \sum_{k=1}^{n} y_k^2 - \hat{\alpha}\sum_{k=1}^{n} y_k - \hat{\beta}\sum_{k=1}^{n} x_k y_k. \tag{18.67}$$

Proof. A straightforward computation yields

$$\begin{aligned}
&\sum_{k=1}^{n}\left(y_k - \left(\hat{\alpha} + \hat{\beta}x_k\right)\right)^2 \\
&= \sum_{k=1}^{n}\left(y_k^2 - 2\left(\hat{\alpha} + \hat{\beta}x_k\right)y_k + \left(\hat{\alpha} + \hat{\beta}x_k\right)^2\right) \\
&= \sum_{k=1}^{n}\left(y_k^2 - 2\hat{\alpha}y_k - 2\hat{\beta}x_k y_k + \hat{\alpha}^2 + 2\hat{\alpha}\hat{\beta}x_k + \hat{\beta}x_k^2\right) \\
&= \sum_{k=1}^{n} y_k^2 - 2\hat{\alpha}\sum_{k=1}^{n} y_k - 2\hat{\beta}\sum_{k=1}^{n} x_k y_k + n\hat{\alpha}^2 + 2\hat{\alpha}\hat{\beta}\sum_{k=1}^{n} x_k + \hat{\beta}^2\sum_{k=1}^{n} x_k^2 \\
&= \sum_{k=1}^{n} y_k^2 - \hat{\alpha}\sum_{k=1}^{n} y_k - \hat{\beta}\sum_{k=1}^{n} x_k y_k + n\hat{\alpha}^2 - \hat{\alpha}\sum_{k=1}^{n} y_k + 2\hat{\alpha}\hat{\beta}\sum_{k=1}^{n} x_k + \hat{\beta}^2\sum_{k=1}^{n} x_k^2 - \hat{\beta}\sum_{k=1}^{n} x_k y_k.
\end{aligned} \tag{18.68}$$

On the other hand, we have

$$n\hat{\alpha}^2 - \hat{\alpha}\sum_{k=1}^{n} y_k = \hat{\alpha}\left(n\left(\bar{y}_n - \hat{\beta}\ \bar{x}_n\right) - \sum_{k=1}^{n} y_k\right) = -n\hat{\alpha}\hat{\beta}\ \bar{x}_n. \tag{18.69}$$

In addition, from the characterization of $\hat{\beta}$ (see Equation 18.45) we can write

$$\hat{\beta}\sum_{k=1}^{n} x_k^2 - \sum_{k=1}^{n} x_k y_k = n\hat{\beta}\bar{x}_n^2 - n\bar{x}_n\bar{y}_n = n\bar{x}_n\left(\hat{\beta}\bar{x}_n - n\bar{y}_n\right) = -n\hat{\alpha}\ \bar{x}_n,$$

which yields

$$\hat{\beta}^2\sum_{k=1}^{n} x_k^2 - \hat{\beta}\sum_{k=1}^{n} x_k y_k = \hat{\beta}\left(\hat{\beta}\sum_{k=1}^{n} x_k^2 - \sum_{k=1}^{n} x_k y_k\right) = -n\hat{\alpha}\hat{\beta}\ \bar{x}_n. \tag{18.70}$$

On account of (18.69) and (18.70), we Then, obtain

$$n\hat{\alpha}^2 - \hat{\alpha}\sum_{k=1}^{n} y_k + 2\hat{\alpha}\hat{\beta}\sum_{k=1}^{n} x_k + \hat{\beta}^2\sum_{k=1}^{n} x_k^2 - \hat{\beta}\sum_{k=1}^{n} x_k y_k = -n\hat{\alpha}\hat{\beta}\ \bar{x}_n + 2n\hat{\alpha}\hat{\beta}\ \bar{x}_n - n\hat{\alpha}\hat{\beta}\ \bar{x}_n = 0.$$

Replacing the latter in (18.68) the desired (18.67) clearly follows. □

The residual sum of squares, RSS, can be interpreted as *the amout of the total variation of the data set* $(y_k)_{k=1}^n$ *which cannot be explained in terms of the variability of the data set* $(x_k)_{k=1}^n$ *through the linear model*. Note that the sum of squared deviations about the regression line is smaller than the sum of squared deviations about any other line. In fact

Remark 1246 *We have*

$$RSS = \min_{(\alpha,\beta)\in\mathbb{R}^2}\left\{SSE\left(\alpha,\beta\right)\right\}. \tag{18.71}$$

Remark 1247 *We have*

$$TSS = ESS + RSS, \tag{18.72}$$

equivalently

$$1 = \frac{ESS}{TSS} + \frac{RSS}{TSS}. \tag{18.73}$$

As a consequence of what presented above, the quantity ESS/TSS [resp. $(100 \cdot ESS/TSS)\,\%$] represents *the proportion* [resp. *percentage*] *of TSS which is explained by the linear regression*, and the quantity RSS/TSS [resp. $(100 \cdot SSR/TSS)\,\%$] represents *the proportion* [resp. *percentage*] *of TSS which cannot be explained by the linear regression*

Definition 1248 *We call the* mean squared residual*, acronym* MSR*, the sample mean of the squared residuals, that is the positive number*

$$MSR \overset{def}{=} \frac{RSS}{n} \equiv \frac{1}{n}\sum_{k=1}^{n}\hat{u}_k^2 \equiv \frac{1}{n}\sum_{k=1}^{n}\left(y_k - \hat{y}_k\right)^2 \equiv \frac{1}{n}\sum_{k=1}^{n}\left(y_k - \left(\hat{\alpha} + \hat{\beta}x_k\right)\right)^2. \tag{18.74}$$

Definition 1249 *We call the* coefficient of determination*, denoted by R^2, the positive number*

$$R^2 \overset{def}{=} \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{k=1}^{n}\left(y_k - \hat{y}_k\right)^2}{\sum_{k=1}^{n}\left(y_k - \bar{y}_n\right)^2}. \tag{18.75}$$

Remark 1250 *We have*

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{k=1}^{n}\left(y_k - \hat{y}_k\right)^2}{\sum_{k=1}^{n}\left(y_k - \bar{y}_n\right)^2}. \tag{18.76}$$

The higher the value of R^2 the more successfully the variation in the data set $(y_k)_{k=1}^n$ can be explained in terms of the linear dependence on the data set $(x_k)_{k=1}^n$ expressed by the regression line.

Definition 1251 *We call the* adjusted coefficient of determination*, denoted by $\tilde{R}^2$, the positive number*

$$\tilde{R}^2 \overset{def}{=} 1 - \frac{RSS/\left(n-c\right)}{TSS/\left(n-1\right)}, \tag{18.77}$$

where c is the number of regression coefficients.

Note that for a simple linear regression $c = 2$ and $\tilde{R}^2 \approx R^2$. The adjusted coefficient of determination $\tilde{R}^2$ will turn out to be a statistic more useful than R^2 when dealing with multiple linear regression.

Remark 1252 *We have*

$$\tilde{R}^2 = 1 - \frac{n-1}{n-c}\left(1 - R^2\right) = 1 - \frac{n-1}{n-c}\frac{\sum_{k=1}^{n}\left(y_k - \hat{y}_k\right)^2}{\sum_{k=1}^{n}\left(y_k - \bar{y}_n\right)^2}. \tag{18.78}$$

Definition 1253 *We call the* goodness of fit coefficient*, denoted by F, the positive number*

$$F \overset{def}{=} \frac{n-c}{c-1}\frac{ESS}{RSS}, \tag{18.79}$$

where c is the number of regression coefficients.

Remark 1254 *We have*

$$F = \frac{n-c}{c-1}\frac{ESS/TSS}{RSS/TSS} = \frac{\dfrac{R^2}{c-1}}{\dfrac{1-R^2}{n-c}}.$$

As we will see in the next section, the goodness of fit coefficient turns out to be the realization of a Fisher-Snedecor distributed statistics which allows an overall evaluation of the linear model by an hypothesis test.

18.4 Statistics of the Univariate Linear Regression

The estimated intercept $\hat{\alpha}$ and slope $\hat{\beta}$ of the regression line are point estimates of the parameters α and β of the linear regression (18.12). To achieve further inferences about α and β, such as confidence intervals or hypothesis tests, it is convenient to introduce suitable estimators for α and β and investigate their distributions. We tackle this task under the assumption that the data set $(x_k)_{k=1}^n$ is observed before the distributions of these estimators are determined. Otherwise saying, we consider the conditional estimators of α and β given that $X_k = x_k$, for every $k = 1, \ldots, n$. In this context, we assume that the random variables constituting the random sample $(Y_k)_{k=1}^n$ are given by

$$Y_k \stackrel{\text{def}}{=} \alpha + \beta x_k + U_k, \quad \forall k = 1, \ldots, n, \tag{18.80}$$

where α and β are the unknown true values of the parameters of the linear regression and $(U_k)_{k=1}^n$ is still a simple random sample of size n drawn from the error term U. Note that, due to (18.80), the random variables $Y_1, \ldots, Y_n$ turn out to be independent and have the same distribution of U.

Definition 1255 *We call the statistic*

$$\hat{\beta}_n = \frac{\sum_{k=1}^n x_k Y_k - n\bar{x}_n \bar{Y}_n}{\sum_{k=1}^n x_k^2 - n\bar{x}_n^2} \qquad [\text{resp. } \hat{\alpha}_n \stackrel{def}{=} \bar{Y}_n - \bar{x}_n \hat{\beta}_n] \tag{18.81}$$

the slope *[resp. the* intercept*]* OLS (conditional) estimator *of the regression line given the data set* $(x_k)_{k=1}^n$.

Lemma 1256 *We have*

$$\hat{\alpha}_n = \sum_{k=1}^n \left(\frac{1}{n} - \frac{(x_k - \bar{x}_n)\,\bar{x}_n}{(n-1)\,s_{X,n}^2} \right) Y_k \qquad \text{and} \qquad \hat{\beta}_n = \frac{1}{n-1} \sum_{k=1}^n \frac{x_k - \bar{x}_n}{s_{X,n}^2} Y_k. \tag{18.82}$$

Proof. On account of the definition of the sample mean $\bar{Y}_n$ and observed sample variance $s_{X,n}^2$, a straightforward computation yields

$$\hat{\beta}_n = \frac{\sum_{k=1}^n x_k Y_k - \bar{x}_n \sum_{k=1}^n Y_k}{\sum_{k=1}^n x_k^2 - n\bar{x}_n^2} = \frac{\sum_{k=1}^n (x_k - \bar{x}_n)}{(n-1)\,s_{X,n}^2} Y_k = \frac{1}{n-1} \sum_{k=1}^n \frac{x_k - \bar{x}_n}{s_{X,n}^2} Y_k.$$

As a consequence, we can write

$$\hat{\alpha}_n = \frac{1}{n} \sum_{k=1}^n Y_k - \frac{\bar{x}_n}{n-1} \sum_{k=1}^n \frac{(x_k - \bar{x}_n)}{s_{X,n}^2} Y_k = \sum_{k=1}^n \left(\frac{1}{n} - \frac{(x_k - \bar{x}_n)\,\bar{x}_n}{(n-1)\,s_{X,n}^2} \right) Y_k.$$

The proof is complete. □

Proposition 1257 *The estimators* $\hat{\alpha}_n$ *and* $\hat{\beta}_n$ *are a linear combination of the entries of the random sample* $(Y_k)_{k=1}^n$.

Proof. It is immediate consequence of Equation (18.82). □

Proposition 1258 *The estimators* $\hat{\alpha}_n$ *and* $\hat{\beta}_n$ *are unbiased.*

Proof. Combining Equations (18.80) and (18.82), we have

$$\begin{aligned}\hat{\beta}_n &= \frac{1}{n-1}\sum_{k=1}^{n}\frac{x_k-\bar{x}_n}{s_{X,n}^2}\left(\alpha+\beta x_k+U_k\right)\\ &= \frac{1}{(n-1)\,s_{X,n}^2}\left(\alpha\sum_{k=1}^{n}\left(x_k-\bar{x}_n\right)+\beta\sum_{k=1}^{n}\left(x_k-\bar{x}_n\right)x_k+\sum_{k=1}^{n}\left(x_k-\bar{x}_n\right)U_k\right).\end{aligned}$$

On the other hand,

$$\sum_{k=1}^{n}\left(x_k-\bar{x}_n\right)=\left(\sum_{k=1}^{n}x_k-\sum_{k=1}^{n}\bar{x}_n\right)=\left(n\bar{x}_n-n\bar{x}_n\right)=0$$

and

$$\sum_{k=1}^{n}\left(x_k-\bar{x}_n\right)x_k=\left(\sum_{k=1}^{n}x_k^2-\bar{x}_n\sum_{k=1}^{n}x_k\right)=\left(\sum_{k=1}^{n}x_k^2-n\bar{x}_n^2\right)=(n-1)\,s_{X,n}^2.$$

Therefore,

$$\hat{\beta}_n=\frac{1}{(n-1)\,s_{X,n}^2}\left((n-1)\,s_{X,n}^2\beta+\sum_{k=1}^{n}\left(x_k-\bar{x}_n\right)U_k\right)=\beta+\sum_{k=1}^{n}\frac{x_k-\bar{x}_n}{(n-1)\,s_{X,n}^2}U_k.$$

As a consequence,

$$\mathbf{E}\left[\hat{\beta}_n\right]=\mathbf{E}\left[\hat{\beta}+\sum_{k=1}^{n}\frac{x_k-\bar{x}_n}{(n-1)\,s_{X,n}^2}U_k\right]=\beta+\sum_{k=1}^{n}\frac{x_k-\bar{x}_n}{(n-1)\,s_{X,n}^2}\mathbf{E}\left[U_k\right]=\beta.$$

This proves that $\hat{\beta}_n$ is unbiased. Now, considering again Equation (18.81), we can write

$$\hat{\alpha}_n=\bar{Y}_n-\bar{x}_n\;\hat{\beta}_n=\frac{1}{n}\sum_{k=1}^{n}\left(\alpha+\beta x_k+U_k\right)-\bar{x}_n\;\hat{\beta}_n=\alpha+\beta\bar{x}_n+\frac{1}{n}\sum_{k=1}^{n}U_k-\bar{x}_n\;\hat{\beta}_n.$$

Hence, on account that $\hat{\beta}_n$ is unbiased, we have

$$\mathbf{E}\left[\hat{\alpha}_n\right]=\mathbf{E}\left[\alpha+\beta\bar{x}_n+\frac{1}{n}\sum_{k=1}^{n}U_k-\bar{x}_n\;\hat{\beta}_n\right]=\alpha+\beta\bar{x}_n+\frac{1}{n}\sum_{k=1}^{n}\mathbf{E}\left[U_k\right]-\bar{x}_n\mathbf{E}\left[\hat{\beta}_n\right]=\alpha.$$

Thus, also $\hat{\alpha}_n$ is unbiased. □

Proposition 1259 *We have*

$$\mathbf{D}^2\left[\hat{\alpha}_n\right]=\frac{\sum_{k=1}^{n}x_k^2}{n\,(n-1)\,s_{X,n}^2}\sigma_U^2\qquad\text{and}\qquad\mathbf{D}^2\left[\hat{\beta}_n\right]=\frac{\sigma_U^2}{(n-1)\,s_{X,n}^2}.$$

Proof. . □

Proposition 1260 *We have*

$$Cov\left(\hat{\alpha}_n,\hat{\beta}_n\right)=-\frac{\bar{x}_n}{(n-1)\,s_{X,n}^2}\sigma_U^2$$

Proof. . □

Theorem 1261 *The estimators $\hat{\alpha}_n$ and $\hat{\beta}_n$ are the minimum variance estimators of the regression coefficients α and β in the class of all linear unbiased estimators of α and β.*

Proof. . □

Definition 1262 *The estimators $\hat{\alpha}_n$ and $\hat{\beta}_n$ are said to be* the best linear unbiased estimators*, acronym* BLUE.

18.5 Univariate Linear Regression with Gaussian Noise

We stress once more that the results presented so far assume no particular form of the distribution of the error term U. However, the assumption that U is Gaussian distributed implies important consequences on the statistical properties of the BLUE estimators.

Theorem 1263 *Under the assumption that U is Gaussian distributed, also $\hat{\beta}_n$ and $\hat{\alpha}_n$ are Gaussian distributed.*

Proof. The random variables $Y_1, \ldots, Y_n$ are independent. The additional assumption of Gaussianity of the error term U implies that $Y_1, \ldots, Y_n$ are Gaussian distributed. On the other hand, by Lemma 1256, the estimators $\hat{\beta}_n$ and $\hat{\alpha}_n$ are linear combination of $Y_1, \ldots, Y_n$. Hence, applying Corollary 847, the claim immediately follows. □

Definition 1264 *We call the statistic*

$$\hat{Y}_k \stackrel{def}{=} \hat{\alpha}_n + \hat{\beta}_n x_k, \quad \forall k = 1, \ldots, n. \tag{18.83}$$

the kth OLS dependent variable (conditional) estimator *given the data set $(x_k)_{k=1}^n$.*

Definition 1265 *We call the statistic*

$$\hat{U}_k \stackrel{def}{=} Y_k - \hat{Y}_k, \quad \forall k = 1, \ldots, n, \tag{18.84}$$

the kth OLS residual *given the data set $(x_k)_{k=1}^n$.*

Note that the kth OLS estimated value $\hat{y}_k$ of the dependent variable [resp. observed residual $\hat{u}_k$] is just an observed value of the kth OLS dependent variable estimator $\hat{Y}_k$ [resp. residual $\hat{U}_k$], for every $k = 1, \ldots, n$.

Theorem 1266 *Under the assumption that U is Gaussian distributed, the kth OLS residual conditional estimator $\hat{U}_k$ is also Gaussian distributed, for every $k = 1, \ldots, n$.*

Proof. Under the assumption that U is Gaussian distributed, the random variables $U_1, \ldots, U_n$ are independent and Gaussian distributed. It follows that the random variables $Y_1, \ldots, Y_n$ are also independent and Gaussian distributed. On the other hand, on account of (18.82), we can write

$$\begin{aligned}
\hat{U}_k &= Y_k - \left(\sum_{\ell=1}^n \left(\frac{1}{n} - \frac{(x_\ell - \bar{x}_n)\,\bar{x}_n}{(n-1)\,s_{X,n}^2} \right) Y_\ell + \frac{x_k}{n-1} \sum_{\ell=1}^n \frac{x_\ell - \bar{x}_n}{s_{X,n}^2} Y_\ell \right) \\
&= \left(1 - \left(\frac{1}{n} - \frac{(x_k - \bar{x}_n)\,\bar{x}_n}{(n-1)\,s_{X,n}^2} + \frac{x_k\,(x_k - \bar{x}_n)}{(n-1)\,s_{X,n}^2} \right) \right) Y_k - \sum_{\substack{\ell=1 \\ \ell \neq k}}^n \left(\frac{1}{n} - \frac{(x_\ell - \bar{x}_n)\,\bar{x}_n}{(n-1)\,s_{X,n}^2} + \frac{x_k\,(x_\ell - \bar{x}_n)}{(n-1)\,s_{X,n}^2} \right) Y_\ell \\
&= \left(1 - \left(\frac{1}{n} + \frac{(x_k - \bar{x}_n)^2}{(n-1)\,s_{X,n}^2} \right) \right) Y_k - \sum_{\substack{\ell=1 \\ \ell \neq k}}^n \left(\frac{1}{n} + \frac{(x_k - \bar{x}_n)\,(x_\ell - \bar{x}_n)}{(n-1)\,s_{X,n}^2} \right) Y_\ell.
\end{aligned}$$

Hence, $\hat{U}_k$ turns out to be a linear combination of independent and Gaussian distributed random variables, for every $k = 1, \ldots, n$. Therefore, the desired result follows from Corollary 847. □

The following Propositions 1267, 1268, and 1269 can be derived by the analogous Propositions 1235, 1236, and 1237, respectivey, just considering that the latter deal with any observed value $\hat{u}_k$ and $\hat{y}_k$ taken by the estimators $\hat{U}_k$ and $\hat{Y}_k$ presented in Definition 1264 and 1265. However, we show how the proofs of Propositions 1235, 1236, and 1237 given in terms of $\hat{u}_k$ and $\hat{y}_k$ work as well in terms of the estimator $\hat{U}_k$ and $\hat{Y}_k$ themselves.

Proposition 1267 *We have*

$$\sum_{k=1}^{n}\hat{U}_k = 0. \tag{18.85}$$

Proof. Combining Equations (18.83) and (18.84), we can write

$$\begin{aligned}
\frac{1}{n}\sum_{k=1}^{n}\hat{U}_k &= \frac{1}{n}\sum_{k=1}^{n}\left(Y_k - \hat{Y}_k\right)\\
&= \frac{1}{n}\sum_{k=1}^{n}\left(Y_k - \left(\hat{\alpha}_n + \hat{\beta}_n x_k\right)\right)\\
&= \frac{1}{n}\sum_{k=1}^{n}Y_k - \frac{1}{n}\sum_{k=1}^{n}\hat{\alpha}_n - \frac{1}{n}\sum_{k=1}^{n}\hat{\beta}_n\ x_k\\
&= \bar{Y}_n - \hat{\alpha}_n - \bar{x}_n\hat{\beta}_n.
\end{aligned}$$

On the other hand, by Definition 1255, it follows

$$\frac{1}{n}\sum_{k=1}^{n}\hat{U}_k = 0,$$

which clearly implies the desired result. □

Proposition 1268 *We have*

$$\frac{1}{n}\sum_{k=1}^{n}\hat{Y}_k = \bar{Y}_n.$$

Proof. Following the proof of Proposition 1236, we can write

$$\frac{1}{n}\sum_{k=1}^{n}\hat{Y}_k - \bar{Y}_n = \frac{1}{n}\sum_{k=1}^{n}\hat{Y}_k - \frac{1}{n}\sum_{k=1}^{n}Y_k = \frac{1}{n}\sum_{k=1}^{n}\left(\hat{Y}_k - Y_k\right) = \frac{1}{n}\sum_{k=1}^{n}\hat{U}_k = 0.$$

which yields the desired result. □

Proposition 1269 *We have*

$$\sum_{k=1}^{n}\hat{U}_k x_k = 0. \tag{18.86}$$

Proof. Following the proof of Proposition 1237, thanks to (18.81), we can write

$$\begin{aligned}
\sum_{k=1}^{n}\hat{U}_k x_k &= \sum_{k=1}^{n}\left(Y_k - \hat{Y}_k\right)x_k\\
&= \sum_{k=1}^{n}\left(Y_k - \left(\hat{\alpha}_n + \hat{\beta}_n x_k\right)\right)x_k\\
&= \sum_{k=1}^{n}x_k Y_k - \hat{\alpha}_n\sum_{k=1}^{n}x_k - \hat{\beta}_n\sum_{k=1}^{n}x_k^2\\
&= \sum_{k=1}^{n}x_k Y_k - n\bar{x}_n\ \hat{\alpha}_n - \hat{\beta}_n\sum_{k=1}^{n}x_k^2\\
&= \sum_{k=1}^{n}x_k Y_k - \left(\bar{Y}_n - \bar{x}_n\ \hat{\beta}_n\right)n\bar{x}_n\ - \hat{\beta}_n\sum_{k=1}^{n}x_k^2\\
&= \sum_{k=1}^{n}x_k Y_k - n\bar{x}_n\bar{Y}_n - \hat{\beta}_n\left(\sum_{k=1}^{n}x_k^2 - n\bar{x}_n^2\right)\\
&= 0.
\end{aligned}$$

as claimed. □

Proposition 1270 *We have*

$$\mathbf{E}\left[\hat{U}_k\right] = 0. \tag{18.87}$$

Proof. On account of the unbiasedness of the conditional estimators $\hat{\alpha}_n$ and $\hat{\beta}_n$, a straightforward computation yields

$$\begin{aligned}\mathbf{E}\left[\hat{U}_k\right] &= \mathbf{E}\left[Y_k - \hat{Y}_k\right] = \mathbf{E}\left[\alpha + \beta x_k + U_k - \left(\hat{\alpha}_n + \hat{\beta}_n x_k\right)\right] \\ &= \alpha + \beta x_k + \mathbf{E}\left[U_k\right] - \left(\mathbf{E}\left[\hat{\alpha}_n\right] - x_k \mathbf{E}\left[\hat{\beta}_n\right]\right) \\ &= 0,\end{aligned}$$

for every $k = 1, \ldots, n$. □

Proposition 1271 *We have*

$$\mathbf{D}^2\left[\hat{U}_k\right] = \left(1 - \frac{1}{n} - \frac{(x_k - \bar{x}_n)^2}{(n-1)\, s_{X,n}^2}\right)\sigma_U^2, \tag{18.88}$$

for every $k = 1, \ldots, n$.

Proof. From the Proof of Therorem 1266 we know that

$$\begin{aligned}\hat{U}_k &= Y_k - \textstyle\sum_{\ell=1}^n \left(\frac{1}{n} + \frac{(x_k - \bar{x}_n)(x_\ell - \bar{x}_n)}{(n-1)\, s_{X,n}^2}\right) Y_\ell \\ &= \left(1 - \left(\frac{1}{n} + \frac{(x_k - \bar{x}_n)^2}{(n-1)\, s_{X,n}^2}\right)\right) Y_k - \textstyle\sum_{\substack{\ell=1 \\ \ell \neq k}}^n \left(\frac{1}{n} + \frac{(x_k - \bar{x}_n)(x_\ell - \bar{x}_n)}{(n-1)\, s_{X,n}^2}\right) Y_\ell\end{aligned}$$

Hence, a straightforward computation yields

$$\begin{aligned}&\mathbf{D}^2\left[\hat{U}_k\right] \\ &= \left(1 - \left(\frac{1}{n} + \frac{(x_k - \bar{x}_n)^2}{(n-1)\, s_{X,n}^2}\right)\right)^2 \mathbf{D}^2\left[Y_k\right] + \textstyle\sum_{\substack{\ell=1 \\ \ell \neq k}}^n \left(\frac{1}{n} + \frac{(x_k - \bar{x}_n)(x_\ell - \bar{x}_n)}{(n-1)\, s_{X,n}^2}\right)^2 \mathbf{D}^2\left[Y_\ell\right] \\ &= \left(\left(1 - \left(\frac{1}{n} + \frac{(x_k - \bar{x}_n)^2}{(n-1)\, s_{X,n}^2}\right)\right)^2 + \textstyle\sum_{\substack{\ell=1 \\ \ell \neq k}}^n \left(\frac{1}{n} + \frac{(x_k - \bar{x}_n)(x_\ell - \bar{x}_n)}{(n-1)\, s_{X,n}^2}\right)^2\right)\sigma_U^2 \\ &= \left(1 - 2\left(\frac{1}{n} + \frac{(x_k - \bar{x}_n)^2}{(n-1)\, s_{X,n}^2}\right) + \left(\frac{1}{n} + \frac{(x_k - \bar{x}_n)^2}{(n-1)\, s_{X,n}^2}\right)^2 + \textstyle\sum_{\substack{\ell=1 \\ \ell \neq k}}^n \left(\frac{1}{n} + \frac{(x_k - \bar{x}_n)(x_\ell - \bar{x}_n)}{(n-1)\, s_{X,n}^2}\right)^2\right)\sigma_U^2.\end{aligned} \tag{18.89}$$

On the other hand,

$$\begin{aligned}
&\left(\frac{1}{n}+\frac{(x_k-\bar{x}_n)^2}{(n-1)\,s_{X,n}^2}\right)^2+\sum_{\substack{\ell=1\\ \ell\neq k}}^{n}\left(\frac{1}{n}+\frac{(x_k-\bar{x}_n)\,(x_\ell-\bar{x}_n)}{(n-1)\,s_{X,n}^2}\right)^2\\
&=\sum_{\ell=1}^{n}\left(\frac{1}{n}+\frac{(x_k-\bar{x}_n)\,(x_\ell-\bar{x}_n)}{(n-1)\,s_{X,n}^2}\right)^2\\
&=\sum_{\ell=1}^{n}\left(\frac{1}{n^2}+2\frac{(x_k-\bar{x}_n)\,(x_\ell-\bar{x}_n)}{n\,(n-1)\,s_{X,n}^2}+\frac{(x_k-\bar{x}_n)^2\,(x_\ell-\bar{x}_n)^2}{(n-1)^2\,{s_{X,n}^2}^2}\right)\\
&=\frac{1}{n}+\frac{2\,(x_k-\bar{x}_n)}{n\,(n-1)\,s_{X,n}^2}\sum_{\ell=1}^{n}(x_\ell-\bar{x}_n)+\frac{(x_k-\bar{x}_n)^2}{(n-1)\,{s_{X,n}^2}^2}\frac{1}{(n-1)}\sum_{\ell=1}^{n}(x_\ell-\bar{x}_n)^2\\
&=\frac{1}{n}+\frac{2\,(x_k-\bar{x}_n)}{n\,(n-1)\,s_{X,n}^2}\,(n\bar{x}_n-n\bar{x}_n)+\frac{(x_k-\bar{x}_n)^2}{(n-1)\,{s_{X,n}^2}^2}s_{X,n}^2\\
&=\frac{1}{n}+\frac{(x_k-\bar{x}_n)^2}{(n-1)\,s_{X,n}^2}
\end{aligned}\tag{18.90}$$

Combining (18.89) and (18.90), we obtain

$$\mathbf{D}^2\left[\hat{U}_k\right]=\left(1-2\left(\frac{1}{n}+\frac{(x_k-\bar{x}_n)^2}{(n-1)\,s_{X,n}^2}\right)+\frac{1}{n}+\frac{(x_k-\bar{x}_n)^2}{(n-1)\,s_{X,n}^2}\right)\sigma_U^2=\left(1-\frac{1}{n}-\frac{(x_k-\bar{x}_n)^2}{(n-1)\,s_{X,n}^2}\right)\sigma_U^2,$$

as desired. □

Proposition 1272 *Under the assumption that U is Gaussian distributed, we have*

$$\frac{1}{\sigma_U^2}\sum_{k=1}^{n}U_k^2\sim\chi_n^2\tag{18.91}$$

and

$$\frac{1}{\sigma_U^2}\sum_{k=1}^{n}\hat{U}_k^2\sim\chi_{n-2}^2.\tag{18.92}$$

Definition 1273 *We call the* OLS conditional estimator of the error variance $\sigma_U^2\equiv\mathbf{D}^2[U]$, *given the data set $(x_k)_{k=1}^n$, the statistic*

$$S_{U,n}^2\overset{def}{=}\frac{1}{n-2}\sum_{k=1}^{n}\hat{U}_k^2=\frac{1}{n-2}\sum_{k=1}^{n}\left(Y_k-\left(\hat{\alpha}_n+\hat{\beta}_n x_k\right)\right)^2.\tag{18.93}$$

We call the OLS conditional estimate of the error variance σ_U^2, *given the data set $(x_k)_{k=1}^n$, the observed value of the estimator $S_{U,n}^2$, that is the positive number*

$$s_{U,n}^2\equiv\hat{\sigma}_U^2\equiv\frac{RSS}{n-2}=\frac{n}{n-2}MSE.\tag{18.94}$$

We call the OLS conditional estimate of the error standard deviation σ_U, *given the data set $(x_k)_{k=1}^n$, the observed value of the estimator $S_{U,n}$, that is the positive number*

$$s_{U,n}\equiv\hat{\sigma}_U\equiv RSE\equiv\sqrt{\frac{RSS}{n-2}}.$$

Note that the divisor $n-2$ in $\hat{\sigma}_U^2$ is due to the presence of the estimators $\hat{\alpha}_n$ and $\hat{\beta}_n$ for α and β in (18.93) and is motivated by obtaining an unbiased estimator for σ_U^2.

Proposition 1274 *The estimator $S_{U,n}^2$ is unbiased for σ_U^2, that is*

$$\mathbf{E}\left[S_{U,n}^2\right] = \sigma_U^2. \tag{18.95}$$

Proof. Thanks to the linearity of the expectation operator and on account of Equation (18.87), we can write

$$\mathbf{E}\left[S_{U,n}^2\right] = \frac{1}{n-2}\sum_{k=1}^{n}\mathbf{E}\left[\hat{U}_k^2\right] = \frac{1}{n-2}\sum_{k=1}^{n}\mathbf{D}^2\left[\hat{U}_k\right]. \tag{18.96}$$

On the other hand, thanks to Equation (18.88), we have

$$\sum_{k=1}^{n}\mathbf{D}^2\left[\hat{U}_k\right] = \sum_{k=1}^{n}\left(1-\frac{1}{n}-\frac{(x_k-\bar{x}_n)^2}{(n-1)\,s_{X,n}^2}\right)\sigma_U^2 = \left(n-1-\frac{1}{s_{X,n}^2}\frac{1}{(n-1)}\sum_{k=1}^{n}(x_k-\bar{x}_n)^2\right)\sigma_U^2 = (n-2) \tag{18.97}$$

Combining (18.96) and (18.97) Equation (18.95) immediately follows.

Alternatively, thanks to Equation (18.92) in Proposition 1272 we can write

$$\mathbf{E}\left[S_{U,n}^2\right] = \frac{1}{n-2}\mathbf{E}\left[\sum_{k=1}^{n}\hat{U}_k^2\right] = \frac{1}{n-2}\mathbf{E}\left[\sigma_U^2\chi_{n-1}^2\right] = \frac{\sigma_U^2}{n-2}\mathbf{E}\left[\chi_{n-2}^2\right] = \sigma_U^2.$$

□

We know

$$Y_k \overset{\text{def}}{=} \alpha+\beta x_k+U_k, \quad \forall k=1,\ldots,n.$$

$$\hat{Y}_k \overset{\text{def}}{=} \hat{\alpha}_n+\hat{\beta}_n x_k, \quad \forall k=1,\ldots,n.$$

$$\hat{U}_k \overset{\text{def}}{=} Y_k-\hat{Y}_k, \quad \forall k=1,\ldots,n.$$

$$\mathbf{E}[U_k]=\mathbf{E}[U]=0, \qquad \mathbf{D}^2[U_k]=\mathbf{D}^2[U]\equiv\sigma_U^2,$$

$$Skew(U_k)=Skew(U)=\frac{\mathbf{E}\left[(U-\mathbf{E}[U])^3\right]}{\mathbf{D}[U]^3}=\frac{\mathbf{E}[U^3]}{\mathbf{D}[U]^3},$$

$$Kurt(U_k)=Kurt(U)=\frac{\mathbf{E}\left[(U-\mathbf{E}[U])^4\right]}{\mathbf{D}[U]^4}=\frac{\mathbf{E}[U^4]}{\mathbf{D}[U]^4},$$

for every $k=1,\ldots,n$.

$$\mathbf{E}[Y_k]=\mathbf{E}[\alpha+\beta x_k+U_k]=\alpha+\beta x_k+\mathbf{E}[U_k]=\alpha+\beta x_k,$$

$$\mathbf{D}^2[Y_k]=\mathbf{D}^2[\alpha+\beta x_k+U_k]=\mathbf{D}^2[U_k]=\sigma_U^2,$$

$$Skew(Y_k)=\frac{\mathbf{E}\left[(Y_k-\mathbf{E}[Y_k])^3\right]}{\mathbf{D}[Y_k]^3}=\frac{\mathbf{E}[U^3]}{\sigma_U^3}=Skew(U)$$

$$Kurt(Y_k)=\frac{\mathbf{E}\left[(Y_k-\mathbf{E}[Y_k])^4\right]}{\mathbf{D}[Y_k]^4}=\frac{\mathbf{E}[U^4]}{\sigma_U^4}=Kurt(U)$$

$$\mathbf{E}\left[\hat{Y}_k\right]=\mathbf{E}\left[\hat{\alpha}_n+\hat{\beta}_n x_k\right]=\mathbf{E}[\hat{\alpha}_n]+x_k\mathbf{E}\left[\hat{\beta}_n\right]=\alpha+\beta x_k,$$

$$\begin{aligned}\mathbf{D}^2\left[\hat{Y}_k\right] &= \mathbf{D}^2\left[\hat{Y}_k\right] = \mathbf{D}^2\left[\hat{\alpha}_n + \hat{\beta}_n x_k\right] \\ &= \mathbf{D}^2\left[\hat{\alpha}_n\right] + x_k^2\mathbf{D}^2\left[\hat{\beta}_n\right] + 2x_k Cov\left(\hat{\alpha}_n, \hat{\beta}_n\right) \\ &= \frac{\sum_{j=1}^n x_j^2}{n\left(n-1\right)s_{X,n}^2}\sigma_U^2 + \frac{x_k^2}{\left(n-1\right)s_{X,n}^2}\sigma_U^2 - 2\frac{\bar{x}_n x_k}{\left(n-1\right)s_{X,n}^2}\sigma_U^2 \\ &= \left(\frac{1}{n}\sum_{j=1}^n x_j^2 + x_k^2 - 2\bar{x}_n x_k\right)\frac{\sigma_U^2}{\left(n-1\right)s_{X,n}^2}\end{aligned}$$

for every $k = 1, \ldots, n$.

$$\mathbf{E}\left[\hat{U}_k\right] = 0, \qquad \mathbf{D}^2\left[\hat{U}_k\right] = \left(1 - \frac{1}{n} - \frac{\left(x_k - \bar{x}_n\right)^2}{\left(n-1\right)s_{X,n}^2}\right)\sigma_U^2$$

As consequence,

$$\mathbf{E}\left[\hat{U}_k^2\right] = \left(1 - \frac{1}{n} - \frac{\left(x_k - \bar{x}_n\right)^2}{\left(n-1\right)s_{X,n}^2}\right)\sigma_U^2$$

$$\begin{aligned}&\mathbf{E}\left[\left(\hat{U}_k - \mathbf{E}\left[\hat{U}_k\right]\right)^3\right] \\ &= \mathbf{E}\left[\left(\left(1 - \left(\frac{1}{n} + \frac{\left(x_k - \bar{x}_n\right)^2}{\left(n-1\right)s_{X,n}^2}\right)\right)Y_k - \sum_{\substack{\ell=1 \\ \ell\neq k}}^n\left(\frac{1}{n} + \frac{\left(x_k - \bar{x}_n\right)\left(x_\ell - \bar{x}_n\right)}{\left(n-1\right)s_{X,n}^2}\right)Y_\ell - \mathbf{E}\left[\hat{U}_k\right]\right)^3\right].\end{aligned}$$

$$\sum_{k=1}^n\left(Y_k - \left(\hat{\alpha}_n + \hat{\beta}_n x_k\right)\right)^3$$

18.5.1 Confidence Intervals for Regression Parameters

Theorem 1275 *Writing* $S_{U,n} \equiv \sqrt{S_{U,n}^2}$, $s_{X,n} \equiv \sqrt{s_{X,n}^2}$, $S_{\hat{\alpha}_n} \equiv \sqrt{\frac{\sum_{k=1}^n x_k^2}{n}}\frac{S_{U,n}}{s_{X,n}}$, *and* $S_{\hat{\beta}_n} \equiv S_{U,n}/s_{X,n}$, *assuming that* U *is normally distributed, the statistics*

$$\frac{\hat{\alpha}_n - \alpha}{S_{\hat{\alpha}_n}} \quad \textit{and} \quad \frac{\hat{\beta}_n - \beta}{S_{\hat{\beta}_n}},$$

have the Student t*-distribution with* $n-2$ *degree of freedom.*

Proof. On account of Propositions (1257) and (1258), we can write

$$\frac{\hat{\alpha}_n - \alpha}{S_{\hat{\alpha}_n}} = \frac{\hat{\alpha}_n - \alpha}{\sqrt{\frac{\sum_{k=1}^n x_k^2}{n}}\frac{S_n(U)}{s_n(x)}} = \frac{\frac{\hat{\alpha}_n - \alpha}{\sqrt{\frac{\sum_{k=1}^n x_k^2}{n}}\sigma_U/s_n(x)}}{\sqrt{\frac{(n-2)S_{U,n}^2/\sigma_U^2}{(n-2)}}} = \frac{\frac{\hat{\alpha}_n - \mathbf{E}[\alpha_n]}{\mathbf{D}[\alpha_n]}}{\sqrt{\frac{(n-2)S_{U,n}^2/\sigma_U^2}{(n-2)}}}$$

and

$$\frac{\hat{\beta}_n - \beta}{S_{\hat{\beta}_n}} = \frac{\hat{\beta}_n - \beta}{S_n\left(U\right)/s_n\left(X\right)} = \frac{\frac{\hat{\beta}_n - \beta}{\sigma_U/S_n(x)}}{\sqrt{\frac{(n-2)S_{U,n}^2/\sigma_U^2}{(n-2)}}} = \frac{\frac{\hat{\beta}_n - \mathbf{E}\left[\hat{\beta}_n\right]}{\mathbf{D}\left[\hat{\beta}_n\right]}}{\sqrt{\frac{(n-2)S_{U,n}^2/\sigma_U^2}{(n-2)}}}.$$

Now, under the assumption that U is normally distributed, by virtue of Proposition (1259), the standardized variables $(\hat{\alpha}_n - \mathbf{E}[\alpha_n])/\mathbf{D}[\alpha_n]$ and $\left(\hat{\beta}_n - \mathbf{E}\left[\hat{\beta}_n\right]\right)/\mathbf{D}\left[\hat{\beta}_n\right]$ are normally distributed. In addition, $(n-2)\,S_{U,n}^2/\sigma_U^2 \sim \chi_{n-2}^2$. It is also possible to prove that both $\hat{\alpha}_n$ and $\hat{\beta}_n$ are independent of $S_{U,n}^2$. Hence, the statistics $(\hat{\alpha}_n - \alpha)/S_{\hat{\alpha}_n}$ and $\left(\hat{\beta}_n - \beta\right)/S_{\hat{\beta}_n}$ turn out to be standard normal random variables divided the square root of an independent chi-square random variable with $n-2$ degree of freedom. This gives the desired result. □

Proposition 1276 *Given any $\varepsilon \in (0,1)$, the $100\,(1-\varepsilon)\,\%$ confidence interval for the intercept parameter α and the slope parameter β are given by*

$$\left(\hat{\alpha} - t_{\varepsilon/2,n-2}s_{\hat{\alpha}_n},\ \hat{\alpha} - t_{\varepsilon/2,n-2}s_{\hat{\alpha}_n}\right) \tag{18.98}$$

and

$$\left(\hat{\beta} - t_{\varepsilon/2,n-2}s_{\hat{\beta}_n},\ \hat{\beta} - t_{\varepsilon/2,n-2}s_{\hat{\beta}_n}\right) \tag{18.99}$$

where $\hat{\alpha}$ [resp. $\hat{\beta}$] is the observed value of the estimator $\hat{\alpha}_n$ [resp. $\hat{\beta}_n$], $s_{\hat{\alpha}_n}$ [resp. $s_{\hat{\beta}_n}$] is the observed value of the estimator $S_{\hat{\alpha}_n}$ [resp. $S_{\hat{\beta}_n}$] of the standard deviation of $\hat{\alpha}_n$ [resp. $\hat{\beta}_n$], and $t_{\varepsilon/2,n-2}$ is the $\varepsilon/2$-critical value of the Student t-distribution with $n-2$ degree of freedom.

Proof. . □

Note that, as usual, the $100\,(1-\varepsilon)\,\%$ confidence interval for the intercept [resp. slope] parameter α [resp. β] is centered at the point estimate $\hat{\alpha}$ [resp. $\hat{\beta}$] of α [resp. β] via the estimator $\hat{\alpha}_n$ [resp. $\hat{\beta}_n$] and its width depends on both on the desired confidence level ε and the observed value $s_{\hat{\alpha}_n}$ [resp. $s_{\hat{\beta}_n}$] of the estimator $S_{\hat{\alpha}_n}$ [resp. $S_{\hat{\beta}_n}$] of the standard deviation of $\hat{\alpha}_n$ [resp. $\hat{\beta}_n$].

18.5.2 Hypothesis Testing for Regression Parameters

The null hypothesis test about β has the form $H_0 : \beta = \beta_0$. In this case the test statistic takes the form

$$\frac{\hat{\beta}_n - \beta_0}{S_{\hat{\beta}_n}}.$$

Under the assumption that H_0 is true, the statistic turns out to be standardized and has the Student t-distribution with $n-2$ dgrees of freedom. As a consequence, given any $\varepsilon \in (0,1)$, the alternative hypotheses and the correspnding rejection regions for the ε-level test are presented in the following table

Alternative Hypothesis	Rejection Region for the ε-Level Test
$H_a : \beta \neq \beta_0$	$t \leq -t_{\varepsilon/2,n-2}$ or $t \geq t_{\varepsilon/2,n-2}$
$H_a : \beta < \beta_0$	$t \leq -t_{\varepsilon,n-2}$
$H_a : \beta > \beta_0$	$t \geq t_{\varepsilon/2,n-2}$

Definition 1277 *We call* model utility t test *the t test under the null hypothesis $H_0 : \beta = 0$ and the alternative $H_a : \beta \neq 0$.*

When this hypothesis is true the knowledge of the data set $(x_k)_{k=1}^n$ gives no information about the variation in the data set $(y_k)_{k=1}^n$. Unless n is quite small, in the model utility test, H_0 will be rejected and the utility of the linear regression model confirmed when r^2 is reasonably large. Unless the model utility test results in a rejection of H_0 for a suitably small ε the linear regression model should not be used for inference and predictions.

18.5.3 Confidence Bands for Estimated Values

Assume U is Gaussian distributed.

Given any $\varepsilon \in (0,1)$, the $100\,(1-\varepsilon)\,\%$ confidence interval (mean response) for the mean $\mathbf{E}\left[\hat{Y}_k\right]$ of the kth OLS dependent variable (conditional) estimator $\hat{Y}_k \equiv \mathbf{E}\left[Y \mid X = x_k\right]$ (corresponding to the observed value x_k of the independent variable X) is given by

$$\left(\hat{y}_k - t_{\frac{\varepsilon}{2},n-2} s_{U,n} \sqrt{\frac{1}{n} + \frac{(x_k - \bar{x}_n)^2}{(n-1)\, s_{X,n}^2}}, \hat{y}_k - t_{\frac{\varepsilon}{2},n-2} s_{U,n} \sqrt{\frac{1}{n} + \frac{(x_k - \bar{x}_n)^2}{(n-1)\, s_{X,n}^2}}\right)$$

for every $k = 1, \ldots, n$, where $\hat{y}_k$ is the realization of $\hat{Y}_k$, that is the kth OLS estimated value of the dependent variable Y. Note that the width of the confidence interval for $\mathbf{E}\left[\hat{Y}_k\right]$ increases as the distance of x_k from the observed sample mean $\bar{x}_n$ increases.

18.5.4 Prediction Bands for Estimated Values

Assume U is Gaussian distributed.

Given any $\varepsilon \in (0,1)$, the $100\,(1-\varepsilon)\,\%$ prediction interval (predicted response) for the value y of the dependent variable Y corresponding to the value x of the dependent variable X is given by

$$\left(\hat{\alpha}_n + \hat{\beta}_n x - t_{\frac{\varepsilon}{2},n-2} s_{U,n} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x}_n)^2}{(n-1)\, s_{X,n}^2}}, \hat{\alpha}_n + \hat{\beta}_n x - t_{\frac{\varepsilon}{2},n-2} s_{U,n} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x}_n)^2}{(n-1)\, s_{X,n}^2}}\right).$$

Note that the width of the confidence interval for y increases as the distance of x from the observed sample mean $\bar{x}_n$ increases.

Part V

Appendices

Chapter 19

Elements of Set Theory

19.1 Preliminaries

The sharp definition of the concepts of *element* and *set* (of elements) has been the subject of deep debates in the Foundations of Mathematics and has undergone a long process of progressive refinement from the "naive" definition given by Georg Cantor. Cantor's definition eventually deals with these concepts as "self-evident", not further specifiable unless resorting to definitional loops. However, for our purposes, Cantor's definition constitutes an acceptable starting point, and we will not dwell on the famous antinomies (see e.g. Bertrand Russel's paradox) it gives rise.

Definition 1278 (Cantor) *A* set *is a gathering together into a whole of definite, distinct objects of our perception or of our thought, which are called* elements *of the set.*

Axiom 1279 *We assume that between an element and a set always occurs one and only one of the two relations:*

1. *the element* belongs *to the set (equivalently, the set* contains *the element);*

2. *the element* does not belong *to the set (equivalently, the set* does not contain *the element).*

Axiom 1280 *We assume that there exists a unique set which contains no elements. We call this set the* empty *set.*

Definition 1281 *We say that a set is* non-empty *if there exists an element in the set.*

Notation 1282 *Generally, will use upper-case letters of the Latin and Greek alphabets to denote sets, e.g. A, B, C, ..., Ω, ..., considering also some special formatting for the Latin letters such as blackboard bold, e.g. $\mathbb{X}$, $\mathbb{Y}$, ... and lower-case letters of the Latin and Greek alphabets for the elements of a set, e.g. a, b, c, ..., ω, ..., x, y.*

Notation 1283 *To denote that an element a belongs [resp. does not belong] to a set A we will write $a \in A$ [resp. $a \notin A$].*

Notation 1284 *Another usual way to denote a set is to write all the elements in the set between braces. For instance, the symbol*

$$\{a, b, c\}$$

will denote the set containing only the first three letters of the Latin alphabet; the symbols

$$\{0, 1\}, \qquad \{0, 1, 2, 3, 4, 5, 9\}, \qquad \{0, 1, 2, 3, 4, 5, 9, A, B, C, D, E, F\}$$

will denote the set of the digits in the binary, decimal and hexadecimal numbering, respectively.

It is important to distinguish between an element, to say a, and the set $\{a\}$ containing the single element a.

Definition 1285 *A set containing a single element is called a* singleton.

Notation 1286 *The symbol* $\{\,\}$ *denotes the empty set. However the empty set is more often denoted by* $\varnothing$.

Notation 1287 *We assume the reader is familiar with the most common sets of numbers, namely: the* natural numbers, *denoted by the symbol* $\mathbb{N}$; *the* integer numbers, *denoted by* $\mathbb{Z}$; *the* rational numbers, $\mathbb{Q}$; *the real numbers,* $\mathbb{R}$; *the complex numbers,* $\mathbb{C}$.

Notation 1288 *In the following we will often use the following symbols:*

1. $=$ *with the meaning of "equal" by virtue of some argument or computation;*
2. $\stackrel{def}{=}$ *with the meaning of "equal" by definition;*
3. $\equiv$ *with the meaning of "equal" by a change of notation;*
4. $\neq$ *with the meaning of "not equal";*
5. $\forall$ *with the meaning of "for every element", "for any element", "for all elements", "given any element";*
6. $\exists$ *with the meaning of "there exists (at least) an element";*
7. $\exists!$ *with the meaning of "there exists a unique element";*
8. $\nexists$ *with the meaning of "there exists no element";*
9. $:$ *with the meaning of "such that";*
10. $\Rightarrow$ *for the logical implication with the meaning of "implies", "if ... then";*
11. $\Leftrightarrow$ *for the logical equivalence with the meaning of "equivalent", "if and only if" by virtue of some argument;*
12. $\stackrel{def}{\Leftrightarrow}$ *with the meaning of "if and only if" by definition;*
13. $\vee$ *for the logical connective* or, *that is the Latin* vel;
14. $\wedge$ *for the logical connective* and, *that is the Latin* et;

15. $\neg$ *for the logical negation* non.

Definition 1289 *We call* predicate *a property concerning the elements of a set which can be unambiguously affirmed or denied. If we denote by* $p(\cdot)$ *a predicate concerning the elements of a set* X *we necessarily have*

$$p(x) = T \Leftrightarrow p(x) \neq F.$$

For instance, with reference to the set of all natural numbers $\mathbb{N}$ the property of being an odd number or an even number or a prime number are predicates. In contrast, the property of being a lucky number is not.

A predicate may result from the combination of other predicates by means of the logical connectives $\vee$ and $\wedge$ or the logical negation $\neg$. For instance, if $p(\cdot)$ and $q(\cdot)$ are predicates concerning the elements of a set X, the notation $p(\cdot) \vee q(\cdot)$ [resp. $p(\cdot) \wedge q(\cdot)$] will represent the predicate which is true for the elements $x \in X$ such that $p(\cdot)$ or $q(\cdot)$ [resp. $p(\cdot)$ and $q(\cdot)$] are true. Formally

$$\begin{aligned} p(x) \vee q(x) = T \quad &\Leftrightarrow \quad p(x) = T \vee q(x) = T, \\ p(x) \wedge q(x) = T \quad &\Leftrightarrow \quad p(x) = T \wedge q(x) = T. \end{aligned}$$

Note that

$$\begin{aligned} p(x) \vee q(x) = F \quad &\Leftrightarrow \quad p(x) = F \wedge q(x) = F, \\ p(x) \wedge q(x) = F \quad &\Leftrightarrow \quad p(x) = F \vee q(x) = F. \end{aligned}$$

The notation $\neg p(\cdot)$ will represent the predicate which is true [false] for the elements $x \in X$ such that if $p(\cdot)$ is false [resp. true]. Formally

$$\neg p(x) = T \quad \Leftrightarrow \quad p(x) = F.$$

Note that the predicate $p(\cdot) \vee \neg p(\cdot)$ is true for any element $x \in X$ and the predicate $p(\cdot) \wedge \neg p(\cdot)$ is false for any element $x \in X$. Formally,

$$\begin{aligned} p(x) \vee \neg p(x) = T, \quad \forall x \in X, \\ p(x) \wedge \neg p(x) = F, \quad \forall x \in X. \end{aligned}$$

Slightly more complicated, the notation $(p(\cdot) \wedge \neg q(\cdot)) \vee (\neg p(\cdot) \wedge q(\cdot))$ will represent the predicate which is true for any element $x \in X$ such that $p(\cdot)$ is true and $q(\cdot)$ is false or $p(\cdot)$ is false and $q(\cdot)$ is true. This is the predicate $p(\cdot)$ *aut* $q(\cdot)$.

19.2 Sets and Subsets

Let $\mathbb{X}$ and $\mathbb{Y}$ be sets.

Definition 1290 *We say that* $\mathbb{X}$ *and* $\mathbb{Y}$ *are equal, and we write* $\mathbb{X} = \mathbb{Y}$*, if they contains the same elements.*

Definition 1291 *We say that* $\mathbb{X}$ *is a* subset *of* $\mathbb{Y}$ *or* $\mathbb{X}$ *is* contained *in* $\mathbb{Y}$ *or* $\mathbb{X}$ *is* included *in* $\mathbb{Y}$*, and we write* $\mathbb{X} \subseteq \mathbb{Y}$*, if any element* $x \in \mathbb{X}$ *belongs to* $\mathbb{Y}$*. In Symbols,*

$$\mathbb{X} \subseteq \mathbb{Y} \quad \overset{def}{\Leftrightarrow} \quad \forall x \in \mathbb{X} \Rightarrow x \in \mathbb{Y}.$$

We say that $\mathbb{X}$ *is a* proper subset *of* $\mathbb{Y}$ *or* $\mathbb{X}$ *is* strictly contained *in* $\mathbb{Y}$ *or* $\mathbb{X}$ *is* strictly included *in* $\mathbb{Y}$*, and we write* $\mathbb{X} \subset \mathbb{Y}$*, if* $\mathbb{X}$ *is a subset of* $\mathbb{Y}$ *and there exists at least an element* $y \in \mathbb{Y}$ *which does not belong to* $\mathbb{X}$*. In Symbols,*

$$\mathbb{X} \subset \mathbb{Y} \quad \overset{def}{\Leftrightarrow} \quad \mathbb{X} \subseteq \mathbb{Y} \ \wedge \ \exists y \in \mathbb{Y} : y \notin \mathbb{X}.$$

Definition 1292 *To indicate that* $\mathbb{X}$ *is a* subset *[resp. a* proper subset*] of* $\mathbb{Y}$ *we also may say that* $\mathbb{Y}$ contains *[resp.* properly contains*]* $\mathbb{X}$ *and write* $\mathbb{Y} \supseteq \mathbb{X}$ *[resp.* $\mathbb{Y} \supset \mathbb{X}$*]. In Symbols*

$$\mathbb{Y} \supseteq \mathbb{X} \quad \textit{[resp. } \mathbb{Y} \supset \mathbb{X}\textit{]} \quad \overset{def}{\Leftrightarrow} \quad \mathbb{X} \subseteq \mathbb{Y} \quad \textit{[resp. } \mathbb{X} \subset \mathbb{Y}\textit{]}.$$

Criterion 1293 (Axiom of Extensionality) *The sets* $\mathbb{X}$ *and* $\mathbb{Y}$ *are equal if and only if* $\mathbb{X}$ *is a* subset *of* $\mathbb{Y}$ *and* $\mathbb{Y}$ *is a* subset *of* $\mathbb{X}$*. Formally,*

$$\mathbb{X} = \mathbb{Y} \quad \Leftrightarrow \quad \mathbb{X} \subseteq \mathbb{Y} \ \wedge \ \mathbb{Y} \subseteq \mathbb{X}.$$

Proposition 1294 *The inclusion relation* $\subseteq$ *fulfills the following properties:*

1. $\mathbb{X} \subseteq \mathbb{X}$ *for any set* $\mathbb{X}$ *(*reflexivity*);*
2. $\mathbb{W} \subseteq \mathbb{X}$ *and* $\mathbb{X} \subseteq \mathbb{Y} \Rightarrow \mathbb{W} \subseteq \mathbb{Y}$ *for all sets* $\mathbb{W}$, $\mathbb{X}$, $\mathbb{Y}$ *(*transitivity*);*
3. $\mathbb{X} \subseteq \mathbb{Y}$ *and* $\mathbb{Y} \subseteq \mathbb{X} \Rightarrow \mathbb{X} = \mathbb{Y}$ *for all sets* $\mathbb{X}$, $\mathbb{Y}$ *(*antisymmetry*).*

Proposition 1295 *The strict inclusion relation* $\subset$ *fulfills the following properties:*

1. $\mathbb{X} \not\subset \mathbb{X}$ *for any set* $\mathbb{X}$ *(*irreflexivity*);*
2. $\mathbb{W} \subset \mathbb{X}$ *and* $\mathbb{X} \subset \mathbb{Y} \Rightarrow \mathbb{W} \subset \mathbb{Y}$ *for all sets* $\mathbb{W}$, $\mathbb{X}$, $\mathbb{Y}$ *(*transitivity*);*
3. $\mathbb{X} \subset \mathbb{Y} \Rightarrow \mathbb{Y} \not\subset \mathbb{X}$ *for all sets* $\mathbb{X}$, $\mathbb{Y}$ *(*asymmetry*).*

Remark 1296 *Given a predicate* $p(\cdot)$ *concerning the elements of* $\mathbb{X}$*, it is natural to consider the subset of the elements of* $\mathbb{X}$ *for which the predicate is true. It is common to denote this set as*

$$\{x \in \mathbb{X} : p(x)\}.$$

According the introduced notation, we clearly have

$$\{x \in \mathbb{X} : p(x) \ \wedge \ \neg p(x)\} = \varnothing \quad \textit{and} \quad \{x \in \mathbb{X} : p(x) \ \vee \ \neg p(x)\} = \mathbb{X}.$$

Example 1297 *The subset* $\mathbb{E} \equiv \{2, 4, ...\}$ *of even natural numbers can denoted by*

$$\{n \in \mathbb{N} : n = 2k, \ k \in \mathbb{N}\}.$$

Similarly, the subset $\mathbb{O} \equiv \{1, 3, 5, ...\}$ *of odd natural numbers can denoted by*

$$\{n \in \mathbb{N} : n = 2k + 1, \ k \in \mathbb{N}\}.$$

19.3 Power Set

Let $\mathbb{X}$ be a set.

Definition 1298 *We call* power set *of* $\mathbb{X}$*, and we denote it by* $\mathcal{P}(\mathbb{X})$*, the* family of all subsets *of* $\mathbb{X}$*. In symbols*

$$\mathcal{P}(\mathbb{X}) \stackrel{def}{=} \{A : A \subseteq \mathbb{X}\}.$$

Remark 1299 *According to Definition 1298, we can write*

$$A \in \mathcal{P}(\mathbb{X}) \Leftrightarrow A \subseteq \mathbb{X}.$$

Example 1300 *Assume* $\mathbb{X} = \{0, 1\}$*. Then,* $\mathcal{P}(\mathbb{X}) = \{\varnothing, \{0\}, \{1\}, \{0, 1\}\}$.

Example 1301 *Assume* $\mathbb{X} = \{a, b, c\}$*. Then,* $\mathcal{P}(\mathbb{X}) = \{\varnothing, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}\}$.

19.4 Operations on a Finite Number of Sets

Let $\mathbb{X}$ be a non-empty set.

Definition 1302 *Given any couple* A*,* B *of subsets of* $\mathbb{X}$*, we call the* union *of* A *and* B*, and we denote it by* $A \cup B$*, the subset of* $\mathbb{X}$ *containing all elements in* A *or* B*. In symbols*

$$A \cup B \stackrel{def}{=} \{x \in \mathbb{X} : x \in A \ \vee \ x \in B\}.$$

Proposition 1303 *Let* A*,* B*,* C *subsets of* $\mathbb{X}$*. We have*

1. $A \cup A = A$;
2. $(A \cup B) \cup C = A \cup (B \cup C)$;
3. $A \cup B = B \cup A$;
4. $A \cup B = B \Leftrightarrow A \subseteq B$*, in particular,* $A \cup \mathbb{X} = \mathbb{X} \Leftrightarrow A = \mathbb{X}$ *and* $A \cup \varnothing = \varnothing \Leftrightarrow A = \varnothing$.

Definition 1304 *Given any couple* A, B *of subsets of* $\mathbb{X}$*, we call the* intersection *of* A *and* B*, and we denote it by* $A \cap B$*, the subset of* $\mathbb{X}$ *containing all elements in* A *or* B*. In symbols*

$$A \cap B \stackrel{def}{=} \{x \in \mathbb{X} : x \in A \ \wedge \ x \in B\}.$$

Proposition 1305 *Let* A*,* B*,* C *subsets of* $\mathbb{X}$*. We have*

1. $A \cap A = A$;
2. $(A \cap B) \cap C = A \cap (B \cap C)$;
3. $A \cap B = B \cap A$;
4. $A \cap B = A \Leftrightarrow A \subseteq B$*, in particular,* $A \cap \mathbb{X} = \mathbb{X} \Leftrightarrow A = \mathbb{X}$ *e* $A \cap \varnothing = A \Leftrightarrow A = \varnothing$.

Definition 1306 *Given any couple A, B of subsets of $\mathbb{X}$, we say that A and B are* disjoint*, if we have*

$$A \cap B = \varnothing.$$

Proposition 1307 (Distributive Laws) *Let A, B, C subsets of $\mathbb{X}$. We have*

1. $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$;
2. $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

Proof. As a simple example of application of the Axiom of Extensionality let us prove that

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C). \tag{19.1}$$

Accordingly, let us try to prove that we have

$$A \cup (B \cap C) \subseteq (A \cup B) \cap (A \cup C) \tag{19.2}$$

and

$$(A \cup B) \cap (A \cup C) \subseteq A \cup (B \cap C). \tag{19.3}$$

Recalling Definition (1291), to show that (19.2) holds true, we will show that every element $x \in A \cup (B \cap C)$ belongs also to $(A \cup B) \cap (A \cup C)$. Conversely, to show that (19.3) holds true, we will show that every element $x \in (A \cup B) \cap (A \cup C)$ belongs also to $A \cup (B \cap C)$.

Let us Then, consider a generic element $x \in A \cup (B \cap C)$. Such an x has to belong to A or $B \cap C$. No other cases are possible. In first case, that is $x \in A$, a fortiori we have $x \in A \cup B$ and $x \in A \cup C$. (in fact both $A \cup B$ and $A \cup C$ contain A). As a consequence, we obtain that $x \in (A \cup B) \cap (A \cup C)$. In the second case, that is $x \in B \cap C$, we have $x \in B$ and $x \in C$. Hence, a fortiori $x \in A \cup B$ and $x \in A \cup C$ ($A \cup B$ and $A \cup C$ contain B and C, respectively). Therefore, we again have $x \in (A \cup B) \cap (A \cup C)$. In both of the two possible cases, we end up with $x \in (A \cup B) \cap (A \cup C)$. This proves (19.2). Now, let us consider a generic element $x \in (A \cup B) \cap (A \cup C)$. Such an x has to belong to $A \cup B$ and $A \cup C$. On the other hand, only two cases are possible: $x \in A$ and $x \notin A$. In the first case, that is $x \in A$, we clearly have $x \in A \cup (B \cap C)$. In the second case, that is $x \notin A$, the belonging of x to both $A \cup B$ and $A \cup C$, clearly implies that x has to belong to both B and C. It Then, follows that $x \in B \cap C$ and a fortiori $x \in A \cup (B \cap C)$. Them, in both of the two possible cases, we end up with $x \in A \cup (B \cap C)$. This proves (19.3). We can finally conclude that (19.1) holds true. □

Definition 1308 *Given any couple A, B of subsets of $\mathbb{X}$, we call the* difference *of A and B, and we denote it by $A - B$, the subset of $\mathbb{X}$ containing all elements in A and not in B. In symbols*

$$A - B \stackrel{def}{=} \{x \in \mathbb{X} : x \in A \ \wedge \ x \notin B\}.$$

Proposition 1309 *Let A, B subsets of $\mathbb{X}$. We have*

1. $A - B = A \Leftrightarrow A \cap B = \varnothing$;
2. $A - B = \varnothing \Leftrightarrow A \subseteq B$.

Definition 1310 *Given any subset A of $\mathbb{X}$, we call the* complement *of A in $\mathbb{X}$, and we denote it by $A^c_{\mathbb{X}}$ the subset of $\mathbb{X}$ containing all elements which do not belong to A. In symbols*

$$A^c_{\mathbb{X}} \stackrel{def}{=} \{x \in \mathbb{X} : x \notin A\}.$$

We will use the shorthand A^c instead of $A^c_{\mathbb{X}}$ when no confusion can arise about $\mathbb{X}$.

Remark 1311 *We clearly have*

$$\mathbb{X}^c_{\mathbb{X}} = \varnothing \quad \textit{and} \quad \varnothing^c_{\mathbb{X}} = \mathbb{X}.$$

Remark 1312 *Let A, B subsets of $\mathbb{X}$. We have*

$$A - B = A \cap B^c_{\mathbb{X}}.$$

Proposition 1313 (De Morgan Laws) *Let A, B subsets of $\mathbb{X}$. We have*

1. $(A \cup B)^c_{\mathbb{X}} = A^c_{\mathbb{X}} \cap B^c_{\mathbb{X}}$;
2. $(A \cap B)^c_{\mathbb{X}} = A^c_{\mathbb{X}} \cup B^c_{\mathbb{X}}$.

Proof. As a another simple example of application of the Axiom of Extensionality let us prove that

$$(A \cup B)^c_{\mathbb{X}} = A^c_{\mathbb{X}} \cap B^c_{\mathbb{X}}. \tag{19.4}$$

Accordingly, let us try to prove that we have

$$(A \cup B)^c_{\mathbb{X}} \subseteq A^c_{\mathbb{X}} \cap B^c_{\mathbb{X}} \tag{19.5}$$

and

$$A^c_{\mathbb{X}} \cap B^c_{\mathbb{X}} \subseteq (A \cup B)^c_{\mathbb{X}}. \tag{19.6}$$

Replicating the argument to prove Proposition 1307, to show that (19.5) holds true, we will show that every element $x \in (A \cup B)^c_{\mathbb{X}}$ belongs also to $A^c_{\mathbb{X}} \cap B^c_{\mathbb{X}}$. Conversely, to show that (19.6) holds true, we will show that every element $x \in A^c_{\mathbb{X}} \cap B^c_{\mathbb{X}}$ belongs also to $(A \cup B)^c_{\mathbb{X}}$.

Let us consider a generic element $x \in (A \cup B)^c_{\mathbb{X}}$. By definition, $x \notin A \cup B$. This happens only if $x \notin A$ and $x \notin B$. Again by definition, it Then, follows that $x \in A^c_{\mathbb{X}}$ and $x \in B^c_{\mathbb{X}}$. Therefore, $x \in A^c_{\mathbb{X}} \cap B^c_{\mathbb{X}}$. This proves (19.5). Conversely, if $x \in A^c_{\mathbb{X}} \cap B^c_{\mathbb{X}}$, Then, $x \in A^c_{\mathbb{X}}$ and $x \in B^c_{\mathbb{X}}$. By definition, $x \notin A$ and $x \notin B$. Hence, $x \notin A \cup B$. Again by definition, $x \in (A \cup B)^c_{\mathbb{X}}$, which proves (19.6). In the end we can conclude that (19.4) holds true. □

19.5 Operation on Families of Sets

Let $\mathcal{F}$ be any non-empty family of subsets of $\mathbb{X}$, that is to say $\varnothing \neq \mathcal{F} \subseteq \mathcal{P}(\mathbb{X})$.

Definition 1314 *We call the* union *of the elements of $\mathcal{F}$, and we denote it by $\bigcup_{F \in \mathcal{F}} F$, the subset of $\mathbb{X}$ containing all elements of $\mathbb{X}$ in at least one of the element of $\mathcal{F}$. In symbols*

$$\bigcup_{F \in \mathcal{F}} F \stackrel{def}{=} \{x \in \mathbb{X} : x \in F,\ \exists F \in \mathcal{F}\}.$$

Definition 1315 *We call the* intersection *of the elements of $\mathcal{F}$, and we denote it by $\bigcap_{F\in\mathcal{F}}F$, the subset of $\mathbb{X}$ containing all elements of $\mathbb{X}$ in all the elements of $\mathcal{F}$. In symbols*

$$\bigcap_{F\in\mathcal{F}}F \stackrel{def}{=} \{x\in\mathbb{X} : x\in F,\ \forall F\in\mathcal{F}\}.$$

Notation 1316 *It is customary to denote a family $\mathcal{F}$ of subsets of $\mathbb{X}$ by means of an indexing set, that is $\mathcal{F}\equiv\{F_j\}_{j\in J}$, where J is a suitable set of indices. In this case, the notation for the union [resp. the intersection] of the elements of $\mathcal{F}$ becomes $\bigcup_{j\in J}F_j$ [resp. $\bigcap_{j\in J}F_j$] with the meaning*

$$\bigcup_{j\in J}F_j \stackrel{def}{=} \{x\in\mathbb{X} : x\in F_j,\ \exists j\in J\} \quad \textit{[resp. } \bigcap_{j\in J}F_j \stackrel{def}{=} \{x\in\mathbb{X} : x\in F_j,\ \forall j\in J\}.$$

19.6 Cartesian Product

Let $\mathbb{X}$, $\mathbb{Y}$ be sets.

Definition 1317 *We call Cartesian product of the sets $\mathbb{X}$ and $\mathbb{Y}$, and we denote it by $\mathbb{X}\times\mathbb{Y}$, the set of all ordered pairs (x,y) having first element $x\in\mathbb{X}$ and second element $y\in\mathbb{Y}$. In symbols*

$$\mathbb{X}\times\mathbb{Y} \stackrel{def}{=} \{(x,y) : x\in\mathbb{X}\ \wedge\ y\in\mathbb{Y}\}.$$

Exercise 1318 *Let $A_1, A_2\subseteq\mathbb{X}$ and $B_1, B_2\subseteq\mathbb{Y}$. We have*

$$(A_1\times B_1)\cap(A_2\times B_2)\subseteq(A_1\cap A_2)\times(B_1\cap B_2).$$

Is the converse true?

Exercise 1319 *Let $A_1, A_2\subseteq\mathbb{X}$ and $B_1, B_2\subseteq\mathbb{Y}$. Does the following equality*

$$(A_1\times B_1)\cup(A_2\times B_2) = (A_1\cup A_2)\times(B_1\cup B_2)$$

hold true?

19.7 Sets and Maps

Let $\mathbb{X}$, $\mathbb{Y}$ be non-empty sets.

Definition 1320 *We call a* map *from $\mathbb{X}$ to $\mathbb{Y}$ any rule which associates with every element $x \in \mathbb{X}$ one and only one element $y \in \mathbb{Y}$. We call the set $\mathbb{X}$ [resp. $\mathbb{Y}$] the* domain *[resp.* codomain*] of the map. We denote a map from $\mathbb{X}$ to $\mathbb{Y}$ by the symbol $f : \mathbb{X} \to \mathbb{Y}$, using the shorthand f when no confusion can arise about the domain $\mathbb{X}$ and the codomain $\mathbb{Y}$.*

Definition 1321 *For each $x \in \mathbb{X}$ we call the f-*image *of x or the* image *of x by f, and we denote it by $f(x)$, the element $y \in \mathbb{Y}$ which is associated with x by the map f.*

Remark 1322 *According to the definition there may exist more than one element of $\mathbb{X}$ having the same image $y \in \mathbb{Y}$ and there may exist some elements of $\mathbb{Y}$ which are images of no element of $\mathbb{X}$.*

Let $f : \mathbb{X} \to \mathbb{Y}$ a map from $\mathbb{X}$ to $\mathbb{Y}$.

Definition 1323 *For any $A \subseteq \mathbb{X}$, we call the f-*image *of A or the* image *of A by f, and we denote it by $f(A)$, the subset of the elements of $\mathbb{Y}$ which are images of some element $x \in A$. In symbols,*

$$f(A) \overset{def}{=} \{y \in \mathbb{Y} : y = f(x), \quad x \in A\}.$$

Let $A_1, A_2 \subseteq \mathbb{X}$.

Proposition 1324 *We have:*

1. $f(A_1 \cup A_2) = f(A_1) \cup f(A_2)$ *for all $A_1, A_2 \subseteq \mathbb{X}$;*
2. $f(A_1 \cap A_2) \subseteq f(A_1) \cap f(A_2)$ *for all $A_1, A_2 \subseteq \mathbb{X}$.*

Remark 1325 *Given $A \subseteq \mathbb{X}$, in general we have neither $f(A^c_{\mathbb{X}}) \subseteq f(A)^c_{\mathbb{Y}}$ nor $f(A)^c_{\mathbb{Y}} \subseteq f(A^c_{\mathbb{X}})$.*

Let

Definition 1326 *For any $B \subseteq \mathbb{Y}$, we call the f-*inverse image *of B or the* inverse image *of B by f, and we denote it by $f^{-1}(B)$, the subset of the elements of A having images in B. In symbols,*

$$f^{-1}(B) \overset{def}{=} \{x \in \mathbb{X} : f(x) \in B\}.$$

Proposition 1327 *We have:*

1. $f^{-1}(B_1 \cup B_2) = f^{-1}(B_1) \cup f^{-1}(B_2)$ *for all $B_1, B_2 \subseteq \mathbb{Y}$;*
2. $f^{-1}(B_1 \cap B_2) = f^{-1}(B_1) \cap f^{-1}(B_2)$ *for all $B_1, B_2 \subseteq \mathbb{Y}$;*
3. $f^{-1}(B^c_{\mathbb{Y}}) = f^{-1}(B)^c_{\mathbb{X}}$ *for every $B \subseteq \mathbb{Y}$.*

Definition 1328 *We call* graph *of f and we denote if by Γ_f the subset of the Cartesian product $\mathbb{X} \times \mathbb{Y}$ given by*

$$\Gamma_f \overset{def}{=} \{(x, y) \in \mathbb{X} \times \mathbb{Y} \mid y = f(x)\}.$$

Definition 1329 *We say that a map $f : \mathbb{X} \to \mathbb{Y}$ is an* injection *or it is* injective *or* one-to-one *if different elements of the domain have different images. In Symbols,*

$$x_1 \neq x_2 \Rightarrow f(x_1) \neq f(x_2), \quad \forall x_1, x_2 \in \mathbb{X}.$$

Remark 1330 *A map $f : \mathbb{X} \to \mathbb{Y}$ is an* injection *if and only if*

$$f(x_1) = f(x_2) \Rightarrow x_1 = x_2, \quad \forall x_1, x_2 \in \mathbb{X}.$$

Remark 1331 *A map $f : \mathbb{X} \to \mathbb{Y}$ is an* injection *if and only if the equation*

$$y = f(x)$$

has at most a solution $x \in \mathbb{X}$ for every $y \in \mathbb{Y}$.

Remark 1332 *A map $f : \mathbb{X} \to \mathbb{Y}$ is an* injection *if and only if the set*

$$\mathbb{X} \times \{y\} \cap \Gamma_f$$

contains at most one point for every $y \in \mathbb{Y}$.

Proposition 1333 *If a map $f : \mathbb{X} \to \mathbb{Y}$ is an* injection *we have*

1. *$f(A_1 \cap A_2) = f(A_1) \cap f(A_2)$ for all A_1, $A_2 \subseteq \mathbb{X}$;*
2. *$f(A_{\mathbb{X}}^c) \subseteq f(A)_{\mathbb{Y}}^c$ for every $A \subseteq \mathbb{X}$.*

Definition 1334 *We say that a map $f : \mathbb{X} \to \mathbb{Y}$ is a* surjection *or it is* surjective *or* onto *if every element of the codomain is image of at least one element of the domain. In Symbols,*

$$\forall y \in \mathbb{Y} \quad \exists x \in \mathbb{X} : \quad y = f(x).$$

Remark 1335 *A map $f : \mathbb{X} \to \mathbb{Y}$ is a* surjection *if and only if*

$$f(\mathbb{X}) = \mathbb{Y}.$$

Remark 1336 *A map $f : \mathbb{X} \to \mathbb{Y}$ is a* surjection *if and only if the equation*

$$y = f(x)$$

has at least a solution $x \in \mathbb{X}$ for every $y \in \mathbb{Y}$.

Remark 1337 *A map $f : \mathbb{X} \to \mathbb{Y}$ is a* surjection *if and only if the set*

$$\mathbb{X} \times \{y\} \cap \Gamma_f$$

contains at least one point for every $y \in \mathbb{Y}$.

Proposition 1338 *If a map $f : \mathbb{X} \to \mathbb{Y}$ is a* surjection *we have*

$$f(A)_{\mathbb{Y}}^c \subseteq f(A_{\mathbb{X}}^c), \quad \forall A \subseteq \mathbb{X}.$$

Definition 1339 *We say that a map $f:\mathbb{X}\to\mathbb{Y}$ is* bijective *or a* bijection *if it is injective and surjective. In Symbols,*

$$\forall y\in\mathbb{Y}\quad \exists!x\in\mathbb{X}:\quad y=f(x).$$

Remark 1340 *A map $f:\mathbb{X}\to\mathbb{Y}$ is* bijective *if and only if the equation*

$$y=f(x)$$

has one and only one solution $x\in\mathbb{X}$ for every $y\in\mathbb{Y}$.

Remark 1341 *A map $f:\mathbb{X}\to\mathbb{Y}$ is* bijective *if and only if the set*

$$\mathbb{X}\times\{y\}\cap\Gamma_f$$

contains one and only one point for every $y\in\mathbb{Y}$.

Remark 1342 *If a map $f:\mathbb{X}\to\mathbb{Y}$ is* bijective *we have*

$$f\left(A_{\mathbb{X}}^{c}\right)=f\left(A\right)_{\mathbb{Y}}^{c},\quad \forall A\subseteq\mathbb{X}.$$

Let $f:\mathbb{X}\to\mathbb{Y}$ and $g:\mathbb{Y}\to Z$ be maps

Definition 1343 *We call* composition *of f and g and we denote it by $g\circ f:\mathbb{X}\to Z$ the map given by*

$$\left(g\circ f\right)(x)\overset{def}{=}g\left(f(x)\right),\quad \forall x\in\mathbb{X}.$$

Remark 1344 *If $f:\mathbb{X}\to\mathbb{Y}$ and $g:\mathbb{Y}\to Z$ are injective [resp. surjective, bijective], Then, the* composition *$g\circ f:\mathbb{X}\to Z$ is injective [resp. surjective, bijective].*

Let $f:\mathbb{X}\to\mathbb{Y}$ a *bijective* map.

Definition 1345 *We call* inverse *of f and we denote it by $f^{-1}:\mathbb{Y}\to\mathbb{X}$ the map given by*

$$f^{-1}\left(y\right)\overset{def}{=}x:\quad f(x)=y.$$

Remark 1346 *We have*

$$f^{-1}\circ f=id_{\mathbb{X}}\quad and\quad f\circ f^{-1}=id_{\mathbb{Y}},$$

where $id_{\mathbb{X}}:\mathbb{X}\to\mathbb{X}$ [resp. $id_{\mathbb{X}}:\mathbb{Y}\to\mathbb{Y}$] is the identity map on $\mathbb{X}$ [resp. $\mathbb{Y}$] given by

$$id_{\mathbb{X}}\left(x\right)\overset{def}{=}x,\quad \forall x\in\mathbb{X}\quad [resp.\ id_{\mathbb{Y}}\left(y\right)\overset{def}{=}y,\quad \forall y\in\mathbb{Y}].$$

Remark 1347 *We have*

$$\left(f^{-1}\circ f\right)(x)\equiv f^{-1}\left(f\left(x\right)\right)=x,\quad \forall x\in\mathbb{X}\quad and\quad \left(f\circ f^{-1}\right)(x)\equiv f\left(f^{-1}\left(y\right)\right)=y,\quad \forall y\in\mathbb{Y}.$$

Definition 1348 *A bijective map is also called* invertible.

Chapter 20

Elements of Combinatorics

20.1 Finite and Infinite Sets

Let $\mathbb{X}$ be a set and let $\mathbb{N} \equiv \{1, 2, \dots\}$ be the set of all natural numbers.

Definition 1349 *We say* $\mathbb{X}$ *is* finite*, if* $\mathbb{X} = \varnothing$ *or there exists an injection from* $\mathbb{X}$ *to a subset* $\{1, 2, \dots, n\}$ *of* $\mathbb{N}$ *for some* $n \in \mathbb{N}$*.*

Proposition 1350 *If* $\mathbb{X} \neq \varnothing$ *is* finite*, Then, there exists a bijection from* $\mathbb{X}$ *to a subset* $\{1, 2, \dots, n\}$ *of* $\mathbb{N}$ *for a unique* $n \in \mathbb{N}$*.*

Definition 1351 *Assume* $\mathbb{X} \neq \varnothing$ *is finite and let* $n \in \mathbb{N}$ *be the unique natural number such that there exist a a bijection from* $\mathbb{X}$ *to* $\{1, 2, \dots, n\}$ *(see Proposition 1350). Then, we say that* $\mathbb{X}$ contains n elements*, and we write* $\#(\mathbb{X}) = n$*. We also say that the empty set* $\varnothing$ *contains zero elements and we write* $\#(\varnothing) = 0$*.*

Definition 1352 *A set* $\mathbb{X}$ *which is not finite is called* infinite*.*

Proposition 1353 *A set* $\mathbb{X}$ *is infinite if and only if there exists a bijection from* $\mathbb{X}$ *to a proper subset of* $\mathbb{X}$*.*

Corollary 1354 *A set* $\mathbb{X}$ *is finite if and only if there exists no bijection from* $\mathbb{X}$ *to a proper subset of* $\mathbb{X}$*.*

20.2 Counting Finite Sets

20.2.1 Factorial and binomial coefficient

Let $\mathbb{N} \equiv \{1, 2, \dots\}$ be the set of all natural numbers and write $\mathbb{N}_0 \equiv \mathbb{N} \cup \{0\}$

Definition 1355 *For any* $n \in \mathbb{N}_0$*, we call* n-factorial*, and we denote it by* $n!$*, the positive integer given by*

$$n! \overset{def}{=} \begin{cases} 1 & \text{if } n = 0 \\ n\,(n-1)! & \text{if } n > 0 \end{cases}.$$

Definition 1356 *For all $k, n \in \mathbb{N}_0$ such that $k \leq n$, we call* binomial coefficient of n over k, *and we denote it by $\binom{n}{k}$ the positive integer given by*

$$\binom{n}{k} \stackrel{def}{=} \frac{n!}{k!\,(n-k)!}.$$

Remark 1357 *For all $k, n \in \mathbb{N}_0$ such that $k \leq n$ we have*

$$\binom{n}{k} = \binom{n}{n-k}.$$

In particular,

$$\binom{n}{0} = \binom{n}{n} = 1.$$

Remark 1358 *For all $k, n \in \mathbb{N}_0$ such that $k \leq n$ we have*

$$\binom{n}{k} + \binom{n}{k+1} = \binom{n+1}{k+1}.$$

Let $\mathbb{R}$ be the set of all real numbers.

Proposition 1359 (Newton's Binomial Theorem) *For every $n \in \mathbb{N}_0$ and all $x, y \in \mathbb{R}$ we have*

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}.$$

In particular,

$$2^n = \sum_{k=0}^{n} \binom{n}{k}.$$

Proposition 1360 *For all $k, m, n \in \mathbb{N}_0$ such that $k \leq \min\{m, n\}$ we have*

$$\binom{m+n}{k} = \sum_{j=0}^{k} \binom{m}{j} \binom{n}{k-j}.$$

Definition 1361 *For all $m, n \in \mathbb{N}_0$ such that $m \geq 1$ and all $k_1, \ldots, k_m \in \mathbb{N}_0$ such that $k_1 + \cdots + k_m = n$, we call* multinomial coefficient *of n over $k_1, \ldots, k_m$, and we denote it by $\binom{n}{k_1, \ldots, k_m}$ the positive integer given by*

$$\binom{n}{k_1, \ldots, k_m} \stackrel{def}{=} \frac{n!}{k_1! \cdots k_m!}.$$

Proposition 1362 (Multinomial Theorem) *For all $m, n \in \mathbb{N}_0$ such that $m \geq 1$ and all $x_1, \ldots, x_m \in \mathbb{R}$ we have*

$$\left(\sum_{j=1}^{m} x_j \right)^n = \sum_{k_1, \ldots, k_m \in \mathbb{N}_0 :\ k_1 + \cdots + k_m = n} \binom{n}{k_1, \ldots, k_m} x_1^{k_1} \cdots x_m^{k_m}.$$

20.2.2 Permutations of order n

Let $n \in \mathbb{N}$.

Definition 1363 *We call* permutation of order n *any bijection from* $\{1, 2, \ldots, n\}$ *to itself.*

Remark 1364 *Every permutation of order n corresponds to an ordered sample of n balls [resp. cards] which can be drawn, without replacement, from a shaken urn [resp. shuffled deck] containing exactly n distinguishable balls [resp. cards].*

Notation 1365 *We write P_n for the number of all permutations of order n.*

Proposition 1366 *We have*

$$P_n \equiv n!.$$

Example 1367 *Assume there are n (distinguishable) runners in a race. Then, we have P_n different possible arrival orders.*

Example 1368 *Assume we have a deck of n playing card. Each time we shuffle the deck we obtain one of the P_n different possible arrangements of the cards in the deck.*

20.2.3 Permutations of order n and class k

Let $k, n \in \mathbb{N}$ such that $k \leq n$.

Definition 1369 *We call* permutation of order n and class k *any injection from* $\{1, 2, \ldots, k\}$ *to* $\{1, 2, \ldots, n\}$.

Remark 1370 *Every permutation of order n and class k corresponds to an ordered sample of k balls [resp. cards] which can be drawn, without replacements, from a shaken urn [resp. shuffled deck] containing n distinguishable balls [resp. cards].*

Notation 1371 *We write $P_{n,k}$ for the number of all permutations of order n and class k.*

Proposition 1372 *We have*

$$P_{n,k} = n\,(n-1)\cdots(n-k+1) = \frac{n!}{(n-k)!}$$

for all $k, n \in \mathbb{N}$ such that $k \leq n$. Note that

$$P_{n,n} = P_n.$$

Example 1373 *Assume in a competition with n players, k rank medals have to be attributed. Then, we have $P_{n,k}$ different possible medal rankings.*

Example 1374 *Assume in a lottery a ordered sequence of k ball is drawn from a urn containing n numbered balls. Then, we have $P_{n,k}$ different possible draws.*

20.2.4 Combinations of order n and class k

Let $k, n \in \mathbb{N}$ such that $k \leq n$.

Definition 1375 *We call* combination of order n and class k *any increasing map from* $\{1, \ldots, k\}$ *to* $\{1, \ldots, n\}$.

Remark 1376 *Every combination of order n and class k corresponds to a subset of $\{1, 2, \ldots, n\}$ containing k elements.*

Remark 1377 *Every combination of order n and class k corresponds to an unordered sample of k balls [resp. cards] which can be drawn, without replacement, from a shaken urn [resp. shuffled deck] containing exactly n distinguishable balls [resp. cards].*

Notation 1378 *We write $C_{n,k}$ for the number of all combinations of order n and class k.*

Proposition 1379 *We have*

$$C_{n,k} = \frac{n!}{k!\,(n-k)!}.$$

Remark 1380 *The number of all subsets of a set containing n elements is 2^n.*

Example 1381 *Assume in a team of n-members, a squad of k-members have to be selected. Then, we have $C_{n,k}$ different possible choices.*

Example 1382 *Assume a hand of k playing cards is dealt from a shuffled deck of n cards. Then, we have $C_{n,k}$ different possible hands.*

20.2.5 Permutations with repetitions of order n and class k

Let $k, n \in \mathbb{N}$.

Definition 1383 *We call* permutation with repetitions of order n and class k *any map from* $\{1, 2, \ldots, k\}$ *to* $\{1, 2, \ldots, n\}$.

Remark 1384 *Any permutation with repetition of order n and class k corresponds to an ordered sample of k balls [resp. cards] which can be drawn, with replacement, from a shaken urn [resp. shuffled deck] containing exactly n distinguishable balls [resp. cards].*

Remark 1385 *Any permutation with repetition of order n and class k can be identified with an element $(x_1, \ldots, x_k)$ of the cartesian product $\{1, 2, \ldots, n\}^k$.*

Notation 1386 *We write $P_{n,k}^{(r)}$ for the number of the permutations with repetitions of order n and class k.*

Proposition 1387 *We have*

$$P_{n,k}^{(r)} = n^k.$$

20.2.6 Permutations of order n with $k_1, \ldots, k_m$ repetitions

Let $m, n \in \mathbb{N}$ such that $m \leq n$ and let $k_1, \ldots, k_m \in \mathbb{N}$ such that $k_1 + \cdots + k_m = n$.

Definition 1388 *We call* permutation of order n with $k_1, \ldots, k_m$ repetitions *any surjection from* $\{1, 2, \ldots, m\}$ *to* $\{1, 2, \ldots, n\}$ *taking the same values on* $k_1, \ldots, k_m$ *elements.*

Remark 1389 *Every permutation of order* n *with* $k_1, \ldots, k_m$ *repetitions corresponds to an ordered sample of* n *balls [resp. cards] which can be drawn, without replacement, from a shaken urn [resp. shuffled deck] containing exactly* n *balls [resp. cards] of* m *distinguishable types such that the* j*th type of balls [resp. cards] consists exactly of* k_j *undistinguishable balls [resp. cards], for* $j = 1, \ldots, m$.

Notation 1390 *We write* $P^{(r)}_{n,k_1,\ldots,k_m}$ *for the number of all permutations of order* n *with* $k_1, \ldots, k_m$ *repetitions.*

Proposition 1391 *We have*

$$P^{(r)}_{n,k_1,\ldots,k_m} = \frac{n!}{k_1! \cdots k_m!}.$$

20.2.7 Combinations with repetitions of order n and class k

Let $k, n \in \mathbb{N}$.

Definition 1392 *We call* combination with repetitions of order n and class k *any non-decreasing map from* $\{1, \ldots, k\}$ *to* $\{1, \ldots, n\}$.

Remark 1393 *Any combination with repetition of order* n *and class* k *corresponds to an unordered sample of* k *balls [resp. cards] which can be drawn, with replacement, from a shaken urn [resp. shuffled deck] containing exactly* n *distinguishable balls [resp. cards].*

Notation 1394 *We write* $C^{(r)}_{n,k}$ *for the number of the combinations with repetitions of order* n *and class* k.

Proposition 1395 *We have*

$$C^{(r)}_{n,k} = \binom{n+k-1}{k}.$$

Proof. To obtain a non-decreasing function of the set $\{1, \ldots, k\}$ into $\{1, \ldots, n\}$ represent the set $\{1, \ldots, n\}$ by n numbered boxes and make up an urn containing k white balls and $n-1$ black ones. After shaking the urn, draw all balls, one after another, and put them in the boxes according to the following rule. Starting from the first box, if the drawn ball is white put it in the box and continue to put white balls into the same box until a black ball is drawn. When a black ball is drawn pass to the second box and repeat the procedure. This placement mechanism is made possible by the presence of exactly $n-1$ black balls in the urn. Every arrangement obtained in such a way corresponds to some nondecreasing mapping of the set $\{1, \ldots, k\}$ into $\{1, \ldots, n\}$. Precisely, the map which assigns to every $j \in \{1, \ldots, k\}$ the number $l \in \{1, \ldots, n\}$ identifying the box where the jth drawn white ball has been placed. On the other hand, this way of sampling from the urn clearly corresponds to a permutation of order $n+k-1$ with k and $n-1$ repetitions. Therefore we have

$$C^{(r)}_{n,k} = P^{(r)}_{2,k,n-1} = \frac{(n+k-1)!}{k!\,(n-1)!},$$

as desired. □

20.2.8 Combinations of order n with $k_1, \ldots, k_m$ repetitions

Let $m, n \in \mathbb{N}$ such that $m \leq n$ and let $k_1, \ldots, k_m,\ \ell_1, \ldots, \ell_m \in \mathbb{N}$ such that $k_j \leq \ell_j$, for every $j = 1, \ldots, m$ and $\ell_1 + \cdots + \ell_m = n$.

Definition 1396 *We call* combination of order n with $k_1, \ldots, k_m$ repetitions *any non-decreasing map from* $\{1, 2, \ldots, m\}$ *to* $\{1, 2, \ldots, n\}$ *taking the same values on* $k_1, \ldots, k_m$ *elements.*

Remark 1397 *Every combination of order* n *with* $k_1, \ldots, k_m$ *repetitions corresponds to an unordered sample of* $k_1 + \cdots + k_m$ *balls [resp. cards] which can be drawn, without replacement, from a shaken urn [resp. shuffled deck] containing exactly* n *balls [resp. cards] of* m *distinguishable types such that the* j*th type of balls [resp. cards] consists exactly of* ℓ_j *undistinguishable balls [resp. cards], for* $j = 1, \ldots, m$.

Notation 1398 *We write* $C^{(r)}_{n,k_1,\ldots,k_m}$ *for the number of the combinations of order* n *with* $k_1, \ldots, k_m$ *repetitions*

Proposition 1399 *We have*

$$C^{(r)}_{n,k_1,\ldots,k_m} = \binom{\ell_1}{k_1} \cdots \binom{\ell_m}{k_m}.$$

20.2.9 Maxwell-Boltzmann, Bose-Einstein, and Fermi-Dirac statistics

Maxwell-Boltzmann, Bose-Einstein, and Fermi-Dirac Statistics consider the number of ways in which m objects can be placed into n cells, according to whether the objects are distinguishable from each other and the cells can contain more than one or only one object.

Case 1400 (Maxwell-Boltzman Statistics) *Suppose we have m distinguishable balls and n boxes. How many ways are there to distribute the balls in the boxes if each box can contain more than one ball? How many ways are there to distribute the balls in the boxes if a specific box has to contain $\ell \leq m$ balls?*

Proof. Since the balls are distinguishable, let us assume that they are numbered from 1 to m. To place a ball in a box means to choose a box. Moreover, since a box can contain more than a ball, a box can be chosen more than once. The first ball can be placed in any of the n boxes. Hence we have n possible ways of placing the first ball. Once the first ball has been placed in a box, the second ball can be placed again in any of the n boxes. In addition, since the balls are distinguishable, the placement of the first ball in the kth box and the second ball in the ℓth box is different from the placement of the first ball in the ℓth box and the second ball in the kth box, whenever $k \neq \ell$. Therefore, there are n^2 possible ways to place the first two balls. Once the first and second balls have been placed in a box, the third ball can be placed again in any of the n boxes, with the same warning on different placements As a consequence, there are n^3 possible ways to place the first three balls. Formalizing this argument by virtue of the Principle of Induction it follows that the number of ways to place m distinguishable balls and n boxes is $n^m = P^{(r)}_{n,m}$, that is the number of the permutations with repetitions of order n and class m.

With regard to the second question, consider that the specific box has to contain ℓ balls among the available m and these balls can be selected in $\binom{m}{\ell}$ different ways. Moreover, once the selected balls have been placed in the box, $n-1$ boxes are left for the placement of the remaining $m-\ell$ balls. In the end, the number

$$\binom{m}{\ell}(n-1)^{m-\ell}$$

yields the desired number of ways to distribute the balls in the boxes. □

Case 1401 (Bose-Einstein Statistics) *Suppose we have m indistinguishable balls and n boxes. How many ways are there in which the balls can be distributed in the boxes if every box can contain more than one ball?*

Proof. Let us number the boxes by $1, \ldots, n$. Then, we can identify every distribution of the balls in the boxes with a combination with repetitions of the numbers $1, \ldots, n$ by putting as many balls in every numbered box as many times the corresponding number appears in the combination. Therefore, the number

$$\binom{n+m-1}{m}$$

yields the number of possible distributions of the Bose-Einstein Statistics. □

Case 1402 (Fermi-Dirac Statistics) *Suppose we have m indistinguishable balls and n boxes. How many ways are there in which the balls can be distributed in the boxes if every box cannot contain more than one ball?*

Proof. Let us number the boxes by $1, \ldots, n$. Note that the impossibility of placing more than one ball in a box (Pauli Exclusion Principle) implies the condition $m \leq n$ binding the number of balls and box under consideration. Now, placing a ball in a box corresponds to select the box. Hence placing m balls in m different boxes corresponds to select m boxes among n. Hence, the number of distribution of the Fermi-Dirac statistics is clearly given by

$$\binom{n}{m}$$

which is the number of all possible subsets of m elements that we can select from a set of n elements. □

Chapter 21

Random Variables on Measurable Spaces

Let $\mathbb{X}$ be a non-empty set and let $\mathcal{M}$ be a σ-algebra of subsets of $\mathbb{X}$.

Definition 1403 (state space) *Following the standard probabilistic terminology, we call the points $x \in \mathbb{X}$* states*, we call the sets $M \in \mathcal{M}$ measurable, and we call the couple $(\mathbb{X}, \mathcal{M})$ a* state space.

Note that in the context of measure theory or functional analysis the triple $(\Omega, \mathcal{E}, \mathbf{P})$ [resp. couple $(\mathbb{X}, \mathcal{M})$] is termed to as a *measure space* [resp. *measurable space*].

Notation 1404 *For brevity, it is customary to denote a probability space with the single symbol Ω rather that the triple $(\Omega, \mathcal{E}, \mathbf{P})$, when no confusion can arise about the σ-algebra of events $\mathcal{E}$ or the probability $\mathbf{P}$. Similarly, we denote a state space with the single symbol $\mathbb{X}$ rather that the couple $(\mathbb{X}, \mathcal{M})$, when no confusion can arise about the σ-algebra of measurable sets $\mathcal{M}$.*

Let $X : \Omega \to \mathbb{X}$, briefly X, be a map from the probability space Ω to the state space $\mathbb{X}$.

Notation 1405 *Given any $M \in \mathcal{M}$, we will use the standard probability notation $\{X \in M\}$, rather than the set-theory notation $X^{-1}(M)$, as an abbreviation to represent the event $\{\omega \in \Omega : X(\omega) \in M\}$, namely the X-inverse image of M.*

Definition 1406 *We say that the map X is an $(\mathcal{E}, \mathcal{M})$-*random variable *on Ω with states in $\mathbb{X}$, if the X-inverse image of any measurable set $M \in \mathcal{M}$ is an event in $\mathcal{E}$. In symbols*

$$\{X \in M\} \in \mathcal{E}, \quad \forall M \in \mathcal{M}. \tag{21.1}$$

In case the σ-algebra of events $\mathcal{E}$ on Ω and σ-algebra of measurable sets $\mathcal{M}$ on $\mathbb{X}$ are fixed once and for all so that there is no danger of confusion, we will omit mentioning them and speak simply of random variable.

Definition 1407 *If X is an $(\mathcal{E}, \mathcal{M})$-*random variable *from the probability space Ω to the state space $\mathbb{X}$, Then, for any $\omega \in \Omega$ the value $X(\omega) \in \mathbb{X}$, namely the X-image of ω, is also called a* state *or a* realization *of X.*

Random variables are aimed to model quantitative or categorical observations on random phenomena. Such observations may return different states, depending on different outcomes of the random phenomena. Clearly, the results of a observations has to be detectable by an observer in light of her information: the observer has to be able to discriminate whether the result of her measurement takes a certain state or falls within a certain set of states. What makes an ordinary map a random variable is just the possibility of observing the events which allow to discriminate the possible states taken by the map, in light of the available information. We know that the available information on a random phenomenon is represented by the σ-algebra of the events on the sample space. Therefore, Equation (**??**) is well suited to represent the idea of observability of the discriminating events. Note also that the notion of random variable on a probabililty space with values in a state space corresponds faithfully to the notion of *measurable map* from a *measure space* to a *measurable space*. As a consequence, all fundamental results for measurable maps developed within *measure theory* remain valid for random variables.

Definition 1408 *If X and Y are $(\mathcal{E},\mathcal{M})$-*random variables *from the probability space Ω to the state space $\mathbb{X}$, the we say that X and Y are equal, and write $X = Y$, if there exists an event $E_0 \in \mathcal{E}$ such that $\mathbf{P}(E_0) = 0$ and*

$$X(\omega) = Y(\omega) \quad \forall \omega \in E_0^c. \tag{21.2}$$

Let $(\Omega,\mathcal{E},\mathbf{P}) \equiv \Omega$ be a probability space and let $(\mathbb{X},\mathcal{M}) \equiv \mathbb{X}$ be a state space.

Example 1409 *Fixed any $x_0 \in \mathbb{X}$, the map $X : \Omega \to \mathbb{X}$ given by*

$$X(\omega) \overset{def}{=} x_0, \quad \forall \omega \in \Omega, \tag{21.3}$$

is an $(\mathcal{E},\mathcal{M})$-random variable, whatever the σ-algebras $\mathcal{E}$ and $\mathcal{M}$ are.

Discussion. Consider any $M \in \mathcal{M}$. Only two cases are possible: $x_0 \in M$, $x_0 \notin M$. In case $x_0 \in M$, we have

$$\{X \in M\} = \Omega \in \mathcal{E}.$$

In the case $x_0 \notin M$, we have

$$\{X \in M\} = \varnothing \in \mathcal{E}$$

Therefore, no matter how $\mathcal{E}$ and $\mathcal{M}$ are given, X is always a $(\mathcal{E},\mathcal{M})$-random variable. □

Definition 1410 *We call the random variable given by Equation (21.3) the* Dirac random variable *concentrated at x_0. A Dirac random variable is also referred to as* deterministic *or* constant *random variable, in the sense that its value x_0 does not change on varying of the sample point $\omega \in \Omega$.*

Note that the Dirac random variable constitutes a model for a deterministic observation in a probabilistic setting.

Example 1411 *Assume there exist $E_0 \in \mathcal{E}$ such that*

$$0 < \mathbf{P}(E_0) < 1, \tag{21.4}$$

and assume there exist at least two distinct states $x_0, x_1 \in \mathbb{X}$. Then, for any non-trivial σ-algebra $\mathcal{M}$ of $\mathbb{X}$, the map $X : \Omega \to \mathbb{X}$ given by

$$X(\omega) \overset{def}{=} \begin{cases} x_0, & \text{if } \omega \in E_0, \\ x_1, & \text{if } \omega \in E_1, \end{cases} \tag{21.5}$$

where $E_1 \equiv E_0^c$, is an $(\mathcal{E},\mathcal{M})$-random variable.

Discussion. Note that,

$$\{X = x_0\} = E_0, \qquad \{X = x_1\} = E_1.$$

On the other hand, since $\mathcal{M}$ is non-trivial, given any $M \in \mathcal{M}$, four cases are are possible:

$$x_0 \in M \text{ and } x_1 \in M, \qquad x_0 \in M \text{ and } x_1 \notin M, \qquad x_0 \notin M \text{ and } x_1 \in M, \qquad x_0 \notin M \text{ and } x_1 \notin M.$$

Therefore, depending on whether case occurs, we have

$$\{X \in M\} = \Omega, \qquad \{X \in M\} = E_0, \qquad \{X \in M\} = E_1, \qquad \{X \in M\} = \varnothing.$$

In any case, we have $\{X \in M\} \in \mathcal{E}$, which yields the desired result. □

Definition 1412 *We call the random variable given by Equations (21.4) and (21.5) the* Bernoulli random variable *with states x_0, x_1, and success probability $\mathbf{P}(X = x_1)$.*

Note that the Bernoulli random variable constitutes a model for the most basic random observation.

Example 1413 *Assume $\mathcal{E}$ is the complete information on Ω, that is $\mathcal{E} = \mathcal{P}(\Omega)$. Then, independently of the choice of $\mathcal{M}$, any map $X : \Omega \to \mathbb{X}$ is an $(\mathcal{E}, \mathcal{M})$-random variable.*

Example 1414 *Assume $\mathcal{E}$ is the trivial information on Ω, that is $\mathcal{E} = \{\varnothing, \Omega\}$. Then, independently of the choice of $\mathcal{M}$, the only $(\mathcal{E}, \mathcal{M})$-random variables $X : \Omega \to \mathbb{X}$ are the Dirac random variables.*

According to Definition 1406, if $\mathcal{E} \equiv \mathcal{P}(\Omega)$, that is the available information on the random phenomenon is the largest attainable, Then, all maps on Ω have observable realizations, to say are $(\mathcal{E}, \mathcal{M})$-measurable. On the contrary, if $\mathcal{E} \equiv \{\Omega, \varnothing\}$, namely the available information is the smallest attainable, Then, only constant maps on Ω have observable realizations. By changing the available information the class of random variables also changes. This is stressed by the following examples.

Example 1415 *Let $\Omega \equiv \{\omega_1, \ldots, \omega_6\}$ be the sample space of all possible outcomes of the roll of a fair die. Consider as a state space $(\mathbb{R}, \mathcal{B}(\mathbb{R})) \equiv \mathbb{R}$ and let $X : \Omega \to \mathbb{R}$ the map given by*

$$X(\omega_j) \overset{def}{=} \begin{cases} j & \text{if } j \text{ is odd} \\ -j+1 & \text{if } j \text{ is even} \end{cases}, \quad \forall j = 1, \ldots, 6.$$

If $\mathcal{E} \equiv \mathcal{P}(\Omega)$ is the complete information on Ω, Then, X is clearly an $(\mathcal{E}, \mathcal{B}(\mathbb{R}))$-random variable. On the other hand, choosing $\mathcal{E} \equiv \{\varnothing, \Omega, \{\omega_1, \omega_3, \omega_5\}, \{\omega_2, \omega_4, \omega_6\}\}$, Then, X is not an $(\mathcal{E}, \mathcal{B}(\mathbb{R}))$-random variable.

Proof. It is sufficient to observe that if $\mathcal{E} \equiv \{\varnothing, \Omega, \{\omega_1, \omega_3, \omega_5\}, \{\omega_2, \omega_4, \omega_6\}\}$, we have

$$(1/2, 3/2) \in \mathcal{B}(\mathbb{R}) \quad \text{and} \quad \{X \in (1/2, 3/2)\} = \{\omega_1\} \notin \mathcal{E}.$$

This implies that X is not an $(\mathcal{E}, \mathcal{B}(\mathbb{R}))$-random variable. □

Remark 1416 *Let $\mathcal{F}$ be a σ-algebra of events of Ω such that $\mathcal{E} \subseteq \mathcal{F}$. Then, any $(\mathcal{E}, \mathcal{M})$-random variable is an $(\mathcal{F}, \mathcal{M})$-random variable.*

Clearly, also changing the σ-algebra $\mathcal{M}$ of measurable sets of $\mathbb{X}$ affects the class of random variables. However, while it rather often arises the need of changing, more specifically reducing, the σ-algebra $\mathcal{E}$ of events of Ω, it hardly ever happens that we have to change the σ-algebra $\mathcal{M}$.

Remark 1417 *Let $\mathcal{N}$ be a σ-algebra of measurable sets on $\mathbb{X}$ such that $\mathcal{N} \subseteq \mathcal{M}$. Then, any $(\mathcal{E}, \mathcal{M})$-random variable is an $(\mathcal{E}, \mathcal{N})$-random variable.*

It is important to introduce the notion of σ-algebra generated by a random variable.

Proposition 1418 *Given any map $X : \Omega \to \mathbb{X}$, the X-inverse image of $\mathcal{M}$, that is the family of events of Ω given by*

$$\sigma(X) \overset{def}{=} \{E \in \mathcal{P}(\Omega) : E = \{X \in M\},\ M \in \mathcal{M}\}, \tag{21.6}$$

is a σ-algebra of events.

Proof. Clearly $\sigma(X)$ is not empty, since $\varnothing = \{X \in \varnothing\}$ and $\Omega = \{X \in \mathbb{X}\}$ are in $\sigma(X)$. Assume $E \in \sigma(X)$, Then, we have $E = \{X \in M\}$, for some $M \in \mathcal{M}$. On the other hand, we know that

$$E^c \equiv \{X \in M\}^c = \{X \in M^c\},$$

where $M^c \in \mathcal{M}$, since $M \in \mathcal{M}$. It Then, follows that $E^c \in \sigma(X)$. Assume $(E_n)_{n\geq 1}$ is a sequence in $\sigma(X)$, then, for every $n \geq 1$ there exists $M_n \in \mathcal{M}$ such that $E_n = \{X \in M_n\}$. We also know that

$$\bigcup_{n=1}^{\infty} E_n \equiv \bigcup_{n=1}^{\infty} \{X \in M_n\} = \left\{X \in \bigcup_{n=1}^{\infty} M_n\right\},$$

where $\bigcup_{n=1}^{\infty} M_n \in \mathcal{M}$, since $M_n \in \mathcal{M}$ for every $n \geq 1$. This implies that also $\bigcup_{n=1}^{\infty} E_n \in \sigma(X)$ and completes the proof. □

Definition 1419 *We call the σ-algebra defined by Equation (21.6) the* σ-algebra generated by *X. Note that in the context of measure theory or functional analysis the σ-algebra $\sigma(X)$ is more commonly denoted by $X^{-1}(\mathcal{M})$ (see Appendix, Definition **??**, Proposition **??**).*

Despite the simplicity of the proof, the following theorem has important consequences.

Theorem 1420 *Given any σ-algebra of events $\mathcal{F}$ on Ω, the map $X : \Omega \to \mathbb{X}$ is an $(\mathcal{F}, \mathcal{M})$-random variable if and only if*

$$\sigma(X) \subseteq \mathcal{F}. \tag{21.7}$$

Proof. Consider an event $E \in \sigma(X)$, Then, $E = \{X \in M\}$ for some $M \in \mathcal{M}$. Under the assumption that X is an $(\mathcal{F}, \mathcal{M})$-random variable, we have $\{X \in M\} \in \mathcal{F}$. Hence, $\sigma(X) \subseteq \mathcal{F}$. Conversely, assume that $\sigma(X) \subseteq \mathcal{F}$, Then, for every $M \in \mathcal{M}$, the event $\{X \in M\} \in \sigma(X)$ is also in $\mathcal{F}$. That is X is an $(\mathcal{F}, \mathcal{M})$-random variable. □

As a consequence of Theorem 1420, the σ-algebra generated by a random variable represents the smallest amount of information which has to be available to observe all events characterized by possible states of the random variable.

A useful criterion to check whether a map $X : \Omega \to \mathbb{X}$ is a random variable is given by the following proposition

Proposition 1421 *Assume $\mathcal{B}$ is a basis for the σ-algebra $\mathcal{M}$ of the measurable sets of the state space $\mathbb{X}$, that is $\sigma(\mathcal{B}) = \mathcal{M}$. Then, the map $X : \Omega \to \mathbb{X}$ is a $(\mathcal{E}, \mathcal{M})$-random variable if and only the X-inverse image of any $B \in \mathcal{B}$ is an event in $\mathcal{E}$. In symbols*

$$\{X \in B\} \in \mathcal{E}, \quad \forall B \in \mathcal{B}. \tag{21.8}$$

Proof. The necessity of Condition 21.8 being obvious, we are left to prove the sufficiency. To this, let us consider the sub-family $\tilde{\mathcal{M}}$ of $\mathcal{M}$ given by

$$\tilde{\mathcal{M}} \equiv \{M \in \mathcal{M} : \{X \in M\} \in \mathcal{E}\}.$$

When Equation 21.8 holds true, we clearly have

$$\mathcal{B} \subseteq \tilde{\mathcal{M}}.$$

In addition, $\tilde{\mathcal{M}}$ is a σ-algebra[1]. We have then

$$\mathcal{M} = \sigma(\mathcal{B}) \subseteq \sigma\left(\tilde{\mathcal{M}}\right) = \tilde{\mathcal{M}},$$

which implies

$$\tilde{\mathcal{M}} = \mathcal{M}.$$

Hence, by virtue of 21.8, for any $M \in \mathcal{M}$ we have $\{X \in M\} \in \mathcal{E}$ which proves that X is a $(\mathcal{E}, \mathcal{M})$-random variable. □

It is also useful to have some tools to build random variables from random variables.

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ a probability space, let $(\mathbb{X}, \mathcal{M}) \equiv \mathbb{X}$ a state space and let $X : \Omega \to \mathbb{X}$ be a map from Ω to $\mathbb{X}$. Let $\mathbb{Y}$ be a non-empty set and let $\mathcal{N}$ be a σ-algebra of subsets of $\mathbb{Y}$. Consider the measurable space $(\mathbb{Y}, \mathcal{N}) \equiv \mathbb{Y}$ and let $g : \mathbb{X} \to \mathbb{Y}$ be an $(\mathcal{M}, \mathcal{N})$-measurable map.

Proposition 1422 *Let $g \circ X : \Omega \to \mathbb{Y}$ be the map given by*

$$(g \circ X)(\omega) \stackrel{def}{=} g(X(\omega)), \quad \forall \omega \in \Omega.$$

We have

$$\sigma(g \circ X) \subseteq \sigma(X).$$

As a consequence, if $X : \Omega \to \mathbb{X}$ is a random variable, Then, also $g \circ X : \Omega \to \mathbb{Y}$ is a random variable.

---

[1] In fact

1. $\tilde{\mathcal{M}} \neq \varnothing$ since $\mathcal{B} \neq \varnothing$;
2. if $M \in \tilde{\mathcal{M}}$ then $\{X \in M\} \in \mathcal{E}$ and since $\mathcal{E}$ is a σ-algebra also $\{X \in M\}^c \in \mathcal{E}$. On the other hand, $\{X \in M\}^c = \{X \in M^c\}$. Hence, $M^c \in \tilde{\mathcal{M}}$.
3. if $(M_n)_{n \geq 1}$ is a sequence belonging to $\tilde{\mathcal{M}}$ then $\{X \in M_n\} \in \mathcal{E}$ for every $n \geq 1$ and since $\mathcal{E}$ is a σ-algebra also $\bigcup_{n \geq 1} \{X \in M_n\} \in \mathcal{E}$. On the other hand, $\bigcup_{n \geq 1} \{X \in M_n\} = \left\{ X \in \bigcup_{n \geq 1} M_n \right\}$. Hence, $\bigcup_{n \geq 1} M_n \in \mathcal{E}$.

Proof. . □

Fixed any $n \in \mathbb{N}$, let $X_1 : \Omega \to \mathbb{X}, \ldots, X_n : \Omega \to \mathbb{X}$ be maps from Ω to $\mathbb{X}$. Let $\mathbb{X}^n \equiv \mathsf{X}_{k=1}^n \mathbb{X}$ be the cartesian product of n copies of $\mathbb{X}$ and let $\mathcal{M}^n \equiv \bigotimes_{k=1}^n \mathcal{M}$ be the tensor product σ-algebra of n copies of $\mathcal{M}$. Consider the measurable space $(\mathbb{X}^n, \mathcal{M}^n) \equiv \mathbb{X}^n$, and let $h : \mathbb{X}^n \to \mathbb{Y}$ be an $(\mathcal{M}^n, \mathcal{N})$-measurable map.

Proposition 1423 *Let $h \circ (X_1, \ldots, X_n) : \Omega \to \mathbb{Y}$ be the map given by*

$$(h \circ (X_1, \ldots, X_n))(\omega) \overset{def}{=} h(X_1(\omega), \ldots, X_n(\omega)), \quad \forall \omega \in \Omega.$$

We have

$$\sigma(h \circ (X_1, \ldots, X_n)) \subseteq \sigma(X_1, \ldots, X_n).$$

As a consequence, if $X_1 : \Omega \to \mathbb{X}, \ldots, X_n : \Omega \to \mathbb{X}$ are random variables, Then, also $h \circ (X_1, \ldots, X_n) : \Omega \to \mathbb{Y}$ is a random variable.

Proof. . □

21.0.10 Distribution of a Random Variable

The *distribution* of a random variable X is just a probability which is built on the state space $\mathbb{X}$ combining the probability $\mathbf{P}$ on the sample space and the random variable itself. The notion of distribution of a random variable is a basic tool that streghtens the connection between Probability and Measure theory while throwing a bridge across Statistics.

Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ a probability space, let $(\mathbb{X}, \mathcal{M}) \equiv \mathbb{X}$ a state space and let $X : \Omega \to \mathbb{X}$ be a map from Ω to $\mathbb{X}$.

Proposition 1424 *Given any $(\mathcal{E}, \mathcal{M})$-random variable, $X : \Omega \to \mathbb{X}$, the map $P_X : \mathcal{M} \to \mathbb{R}_+$ given by*

$$P_X(M) \overset{def}{=} \mathbf{P}(X \in M), \quad \forall M \in \mathcal{M}, \tag{21.9}$$

where $\mathbf{P}(X \in M)$ is the standard shorthand for $\mathbf{P}(\{X \in M\})$, is a probability on $\mathbb{X}$.

Proof. We clearly have

$$\{X \in \mathbb{X}\} \equiv \{\omega \in \Omega : X(\omega) \in \mathbb{X}\} = \Omega.$$

In addition, given any sequence $(M_n)_{n \geq 1}$ of pairwise disjoint sets in $\mathcal{M}$, the sequence $(\{X \in M_n\})_{n \geq 1}$ is a sequence of pairwise incompatible events in $\mathcal{E}$ such that

$$\left\{ X \in \bigcup_{n=1}^{\infty} M_n \right\} = \bigcup_{n=1}^{\infty} \{X \in M_n\}.$$

Therefore, according Equation (21.9), we have

$$P_X(\mathbb{X}) = \mathbf{P}(X \in \mathbb{X}) = \mathbf{P}(\Omega) = 1.$$

Moreover,

$$P_X\left(\bigcup_{n=1}^{\infty} M_n\right) = \mathbf{P}\left(X \in \bigcup_{n=1}^{\infty} M_n\right) = \mathbf{P}\left(\bigcup_{n=1}^{\infty} \{X \in M_n\}\right) = \sum_{n=1}^{\infty} \mathbf{P}(X \in M_n) = \sum_{n=1}^{\infty} P_X(M_n).$$

Hence, Properties 1 and 2 are satisfied. This proves the claim. □

Definition 1425 *We call the* distribution *or* law *of X the probability $P_X : \mathcal{M} \to \mathbb{R}_+$, briefly P_X, given by Equation 21.9.*

Example 1426 *Assume X is a Dirac random variable concentrated at a state $x_0 \in \mathbb{X}$ (see Definition 1410). Then, the distribution $P_X : \mathcal{M} \to \mathbb{R}_+$ of X is given by*

$$P_X(M) \stackrel{def}{=} \mathbf{P}(X \in M) = \begin{cases} 1, & \text{if } x_0 \in M, \\ 0, & \text{if } x_0 \notin M. \end{cases} \tag{21.10}$$

Hence, P_X is the Dirac probability on $\mathbb{X}$ concentrated at x_0 (see Definition 263).

Example 1427 *Assume X is a Bernoulli random variable X with states $x_0, x_1 \in \mathbb{X}$ and success probability $p \equiv \mathbf{P}(X = x_1)$ (see Definition 1412). Then, the distribution $P_X : \mathcal{M} \to \mathbb{R}_+$ of X is given by*

$$P_X(M) \stackrel{def}{=} \mathbf{P}(X \in M) = \begin{cases} 1, & \text{if } x_0 \in M \text{ and } x_1 \in M, \\ p, & \text{if } x_0 \in M \text{ and } x_1 \notin M, \\ q, & \text{if } x_0 \notin M \text{ and } x_1 \in M, \\ 0, & \text{if } x_0 \notin M \text{ and } x_1 \notin M. \end{cases}$$

Hence, P_X is the Bernoulli probability on $\mathbb{X}$ (see Example 265).

Random variables and distributions are dual notions. Just two different ways to represent quantitative observations on random phenomena. Roughly speaking, a random variable and its distribution are like the two sides of the same coin. It may be convenient to look upon the one or the other according to different circumstances. For instance, as we will see, the integral calculus dealing with functionals of a random variable on a probability space can be easily transferred to the integral calculus dealing with functionals on the state space equipped with the measure defined as the distribution of the random variable.

Particularly remarkable from the statistical point of view is the following result.

Theorem 1428 (I Inversion Theorem) *Given any probability $P : \mathcal{M} \to \mathbb{R}_+$ on a state space $(\mathbb{X}, \mathcal{M})$, there always exist a probability space $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ and a random variable $X : \Omega \to \mathbb{X}$ such that the distribution $P_X : \mathcal{M} \to \mathbb{R}_+$ of X is $P : \mathcal{M} \to \mathbb{R}_+$.*

Proof. ... □

Proposition 1429 *Assume $X : \Omega \to \mathbb{X}$ and $Y : \Omega \to \mathbb{X}$ are random variables such that $X = Y$ (see Definition 1408). Consider the distributions $P_X : \mathcal{M} \to \mathbb{R}_+$ and $P_Y : \mathcal{M} \to \mathbb{R}_+$ of X and Y, respectively. Then, we have*

$$P_X = P_Y \tag{21.11}$$

Proof. Since the events $\{X = Y\}$ and $\{X \neq Y\}$ are a partition of Ω such that $\mathbf{P}(X = Y) = 1$ and $\mathbf{P}(X \neq Y) = 0$, for any $M \in \mathcal{M}$ we can write

$$\begin{aligned} P_X(M) \stackrel{\text{def}}{=} & \mathbf{P}(X \in M) \\ & = \mathbf{P}(\{X \in M\} \cap (\{X = Y\} \cup \{X \neq Y\})) \\ & = \mathbf{P}(\{X \in M\} \cap \{X = Y\}) + \mathbf{P}(\{X \in M\} \cap \{X \neq Y\}) \\ & = \mathbf{P}(\{X \in M\} \cap \{X = Y\}) \end{aligned} \tag{21.12}$$

On the other hand, we clearly have

$$\{X \in M\} \cap \{X = Y\} = \{Y \in M\} \cap \{X = Y\}.$$

Therefore,

$$\begin{aligned}
&\mathbf{P}(\{X \in M\} \cap \{X = Y\}) \\
&\qquad = \mathbf{P}(\{Y \in M\} \cap \{X = Y\}) \\
&\qquad = \mathbf{P}(\{Y \in M\} \cap \{X = Y\}) + \mathbf{P}(\{Y \in M\} \cap \{X \neq Y\}) \\
&\qquad = \mathbf{P}(\{Y \in M\} \cap (\{X = Y\} \cup \{X \neq Y\})) \\
&\qquad = \mathbf{P}(Y \in M) \overset{\text{def}}{=} P_Y(M). \qquad (21.13)
\end{aligned}$$

Combining (21.12) and (21.13) the claim follows. □

Note that the converse of Proposition 1429 is not true. That is different random variables may have the same distribution. This circumstance is shown in the following example.

Example 1430 *Choose the sample space $\Omega \equiv \{\omega_0, \omega_1\}$, the σ-algebra $\mathcal{E} \equiv \mathcal{P}(\Omega)$ and the naive probability $\mathbf{P} : \mathcal{E} \to \mathbb{R}_+$ given by*

$$\mathbf{P}(\omega_k) \overset{def}{=} \frac{1}{2}, \quad k = 0, 1.$$

Consider the state space $(\mathbb{R}, B(\mathbb{R}))$ and the Bernoulli random variables $X : \Omega \to \mathbb{R}$ and $Y : \Omega \to \mathbb{R}$ given by

$$X(\omega_k) \overset{def}{=} \begin{cases} 0 & \text{if } k = 0 \\ 1 & \text{if } k = 1 \end{cases}, \qquad Y(\omega_k) \overset{def}{=} \begin{cases} 1 & \text{if } k = 0 \\ 0 & \text{if } k = 1 \end{cases}.$$

Write $P_X : B(\mathbb{R}) \to \mathbb{R}_+$ and $P_Y : B(\mathbb{R}) \to \mathbb{R}_+$ for the distribution of X and Y, respectively, where . We have

$$X \neq Y \quad \text{and} \quad P_X = P_Y.$$

Discussion. Since

$$X(\omega_k) \neq Y(\omega_k), \quad \text{and} \quad \mathbf{P}(\omega_k) \neq 0$$

for every $k = 1, 2$, we cannot have $X = Y$. On the other hand, we have

$$P_X(B) = \mathbf{P}(X \in B) = \begin{cases} \mathbf{P}(\Omega) = 1 & \text{if } 0, 1 \in B \\ \mathbf{P}(\omega_0) = \frac{1}{2} & \text{if } 0 \in B \text{ and } 1 \notin B \\ \mathbf{P}(\omega_1) = \frac{1}{2} & \text{if } 0 \notin B \text{ and } 1 \in B \\ \mathbf{P}(\varnothing) = 0 & \text{if } 0, 1 \notin B \end{cases}$$

and

$$P_Y(B) = \mathbf{P}(Y \in B) = \begin{cases} \mathbf{P}(\Omega) = 1 & \text{if } 0, 1 \in B \\ \mathbf{P}(\omega_1) = \frac{1}{2} & \text{if } 0 \in B \text{ and } 1 \notin B \\ \mathbf{P}(\omega_0) = \frac{1}{2} & \text{if } 0 \notin B \text{ and } 1 \in B \\ \mathbf{P}(\varnothing) = 0 & \text{if } 0, 1 \notin B \end{cases}.$$

It follows that

$$P_X(B) = P_Y(B)$$

for every $B \in B(\mathbb{R})$, and this completes the discussion. □

21.0.11 Density of a Random Variable with States in a Measure Space

Assume the state space $(\mathbb{X}, \mathcal{M})$ is equipped with a σ-finite measure $\mu : \mathcal{M} \to \mathbb{R}_+$ and keep on writing $\mathbb{X}$ as a shortand for the measure space $(\mathbb{X}, \mathcal{M}, \mu)$. Let $X : \Omega \to \mathbb{X}$ be a random variable and let $P_X : \mathcal{M} \to \mathbb{R}_+$ be the distribution of X.

Definition 1431 *We say that X is* absolutely continuous *if the distribution P_X of X is* absolutely continuous *with respect to μ*[2]*.*

Theorem 1432 (Radon-Nikodym) *Assume X is absolutely continuous. Then, there exists a function $f \in \mathcal{L}^1(\mathbb{X}; \mathbb{R})$ such that*

$$\mathbf{P}(X \in M) = \int_M f d\mu, \quad \forall M \in \mathcal{M}. \tag{21.14}$$

In addition, if $g : \mathbb{X} \to \mathbb{R}$ is another function fulfilling (21.14) we have

$$g = f, \quad \mu - a.e$$

Definition 1433 *Assuming X absolutely continuous, we call the* density of X*, and we denote it by f_X, the equivalence class of all the functions in $\mathcal{L}^1(\mathbb{X}; \mathbb{R})$ fulfilling (21.14) which are $\mu - a.e.$ equal. We call a* version *of the* density *of X any function $f \in \mathcal{L}^1(\mathbb{X}; \mathbb{R})$ fulfilling Equation (21.14) of Theorem 1432. However, for our purposes, we can neglect the distinction between the density of a random variable and any version of such a density. Therefore, from now on, by the* density *of the random variable X we will mean any version of the density for which we will retain the notation f_X.*

Proposition 1434 *If $X \equiv 1_E$ is the indicator function of some event $E \in \mathcal{E}$, Then,*

$$\mathbf{E}[1_E \mid \mathcal{F}] = \mathbf{P}(E \mid \mathcal{F}). \tag{21.15}$$

Proof. We clearly have

$$\mathbf{P}_{\mathcal{F}}^{1_E}(F) = \int_F 1_E \, d\mathbf{P} = \int_\Omega 1_E 1_F \, d\mathbf{P} = \int_\Omega 1_{E \cap F} \, d\mathbf{P} = \int_{E \cap F} d\mathbf{P} = \mathbf{P}(E \cap F) = \mathbf{P}_{\mathcal{F}}^{E}(E),$$

for every $F \in \mathcal{F}$. The claim immediately follows. □

[2] We recall from Measure Therory that the distribution P_X is absolutely continuous with respect to μ, in symbols $P_X \ll \mu$, if for any $M \in \mathcal{M}$ such that $\mu(M) = 0$ we have $P_X(M) = 0$.

Part VI

References

Bibliography

[1] Baldi, P. (2011). "Calcolo delle probabilità" (Seconda ed.). McGraw-Hill, Milano.

[2] Campbell, P.F., and McCabe, G.P. (1984). Predicting the Success of Freshmen in a Computer Science Major, Communications of the ACM 27(11): 1108-1113 (DOI: 10.1145/1968.358288)

[3] Cramér, H. (1999). "Mathematical Methods of Statistics", Princeton Landmarks in Mathematics and Physics (PMS-9). Princeton University Press, Princeton

[4] Devore, J.L., and Berk, K.N. (2011). "Modern Mathematical Statistics with Application" (2nd ed.). Springer, New York.

[5] Lessi, O. (1993). "Corso di Probabilità". Metria, Padova.

[6] Rao, M.M., and Swift, R.J. (2005). "Probability Theory with Applications" (2nd ed.). Springer, New York.

[7] Ross, S. (2015). "Probabilità e Statistica per l'Ingegneria e le Scienze" (Terza ed.). Apogeo education. Maggioli, Santarcangelo di Romagna (RM).