Lecture Notes on Stochastic Calculus (Part I)

Fabrizio Gelsomino, Olivier Lévêque, EPFL

November 30, 2009

Contents

1 Probability "review"

1.1 σ-fields

In probability, the *fundamental set* Ω describes the set of all possible *outcomes* (or *realizations*) of a given experiment. It might be *any* set, without any particular structure, such as for example $\Omega = \{1, \ldots, 6\}$ representing the outcomes of a die roll, or $\Omega = [0,1]$ representing e.g. the outcomes of a concentration measurement of some chemical product. Notice moreover that the set Ω need not be composed of numbers exclusively. It is e.g. perfectly valid to consider the set $\Omega = \{\text{banana, apple, orange}\}$.

Given a fundamental set Ω, it is important to describe what *information* does one have on the system, namely on the outcomes of the experiment. This notion of information is well captured by the mathematical notion of *σ-field*, which is defined below. Notice that in elementary probability courses, it is generally assumed that the information one has about a system is *complete*, so that it becomes useless to introduce the concept below.

Definition 1.1. Let Ω be a set. A *σ-field* (or *σ-algebra*) on Ω is a collection $\mathcal{F}$ of subsets of Ω (or *events*) satisfying the following three properties or *axioms*:

(i) $\emptyset \in \mathcal{F}$.

(ii) If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$.

(iii) If $(A_n)_{n=1}^{\infty} \subset \mathcal{F}$, then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$. In particular, if $A, B \in \mathcal{F}$, then $A \cup B \in \mathcal{F}$.

The following properties can be further deduced from the above axioms (this is left as an exercise):

(iv) $\Omega \in \mathcal{F}$.

(v) If $(A_n)_{n=1}^{\infty} \subset \mathcal{F}$, then $\bigcap_{n=1}^{\infty} A_n \in \mathcal{F}$. In particular, if $A, B \in \mathcal{F}$, then $A \cap B \in \mathcal{F}$.

(vi) If $A, B \in \mathcal{F}$ and $A \subset B$, then $B \backslash A \in \mathcal{F}$.

Terminology. The pair $(\Omega, \mathcal{F})$ is called a *measurable space* and the events belonging to $\mathcal{F}$ are said to be *$\mathcal{F}$-measurable*, that is, they are the events that one can decide on whether they happened or not, given the information $\mathcal{F}$. In other words, if one knows the information $\mathcal{F}$, then one is able to tell to which events of $\mathcal{F}$ (= subsets of Ω) does the realization of the experiment ω belong.

Example. For a generic set Ω, the following are always σ-fields:

$\mathcal{F}_0 = \{\emptyset, \Omega\}$ (= *trivial* σ-field).
$\mathcal{P}(\Omega) = \{\text{all subsets of } \Omega\}$ (= *complete* σ-field).

Example 1.2. Let $\Omega = \{1, \ldots, 6\}$. The following are σ-fields on Ω:

$\mathcal{F}_1 = \{\emptyset, \{1\}, \{2, \ldots, 6\}, \Omega\}$.
$\mathcal{F}_2 = \{\emptyset, \{1,3,5\}, \{2,4,6\}, \Omega\}$.

Example 1.3. Let $\Omega = [0,1]$ and $I_1, \ldots, I_n$ be a family of disjoint intervals in Ω such that $I_1 \cup \ldots \cup I_n = \Omega$ ($\{I_1, \ldots, I_n\}$ is also called a *partition* of Ω). The following is a σ-field on Ω:

$$\mathcal{F}_3 = \{\emptyset, I_1, \ldots, I_n, I_1 \cup I_2, \ldots, I_1 \cup I_2 \cup I_3, \ldots, \Omega\} \quad \text{(NB: there are } 2^n \text{ events in total in } \mathcal{F}_3\text{).}$$

σ-field generated by a collection of events.

An event carries in general more information than itself. As an example, if one knows whether the result of a die roll is odd (corresponding to the event $\{1,3,5\}$), then one also knows of course whether the result is even (corresponding to the event $\{2,4,6\}$). It is therefore convenient to have a mathematical description of the information generated by a single event, or more generally by a family of events.

Definition 1.4. Let $\mathcal{A} = \{A_i,\ i \in I\}$ be a collection of events, where I need not be a countable set. The *σ-field generated by* $\mathcal{A}$ is the smallest σ-field on Ω containing all the events A_i. It is denoted as $\sigma(\mathcal{A})$.

Example. Let $\Omega = \{1, \dots, 6\}$ (cf. Example 1.2).

Let $\mathcal{A}_1 = \{\{1\}\}$. Then $\sigma(\mathcal{A}_1) = \mathcal{F}_1$.
Let $\mathcal{A}_2 = \{\{1,3,5\}\}$. Then $\sigma(\mathcal{A}_2) = \mathcal{F}_2$.
Let $\mathcal{A} = \{\{1\}, \dots, \{6\}\}$. Then $\sigma(\mathcal{A}) = \mathcal{P}(\Omega)$.

Exercise. Let $\mathcal{A} = \{\{1,2,3\}, \{1,3,5\}\}$. Compute $\sigma(\mathcal{A})$.

Example. Let $\Omega = [0,1]$ and let $\mathcal{A}_3 = \{I_1, \dots, I_n\}$ (cf. Example 1.3). Then $\sigma(\mathcal{A}_3) = \mathcal{F}_3$.

Borel σ-field. Another important example of generated σ-field on $\Omega = [0,1]$ is the following:

$$\mathcal{B}([0,1]) = \sigma(\{\,]a,b[\,:\ a,b \in [0,1],\ a < b\}),$$

is the *Borel σ-field* on $[0,1]$ and elements of $\mathcal{B}([0,1])$ are called the *Borel subsets* of $[0,1]$. As surprising as it may be, $\mathcal{B}([0,1]) \neq \mathcal{P}([0,1])$, which generates some difficulties from the theoretical point of view. Nevertheless, it is quite difficult to construct explicit examples of subsets of $[0,1]$ which are *not* in $\mathcal{B}([0,1])$.

Sub-σ-field.

One may have more or less information about a system. In mathematical terms, this translates into the fact that a σ-field has more or less elements. It is therefore convenient to introduce a (partial) ordering on the ensemble of existing σ-fields, in order to establish a *hierarchy* of information. This notion of hierarchy is important and will come back when we will be studying stochastic processes that evolve in time.

Definition 1.5. Let Ω be a set and $\mathcal{F}$ be a σ-field on Ω. A *sub-σ-field of* $\mathcal{F}$ is a collection $\mathcal{G}$ of events such that:

(i) If $A \in \mathcal{G}$, then $A \in \mathcal{F}$.

(ii) $\mathcal{G}$ is itself a σ-field.

Notation. $\mathcal{G} \subset \mathcal{F}$.

Remark. Let Ω be a generic set. The trivial σ-field $\mathcal{F}_0 = \{\emptyset, \Omega\}$ is a sub-σ-field of any other σ-field on Ω. Likewise, any σ-field on Ω is a sub-σ-field of the complete σ-field $\mathcal{P}(\Omega)$.

Example. Let $\Omega = \{1, \dots, 6\}$ (cf. Example 1.2). Notice that $\mathcal{F}_1$ is *not* a sub-σ-field of $\mathcal{F}_2$ (even though $\{1\} \subset \{1,3,5\}$), nor is $\mathcal{F}_2$ a sub-σ-field of $\mathcal{F}_1$. In general, notice that

1) If $A \in \mathcal{G}$ and $\mathcal{G} \subset \mathcal{F}$, then it is true that $A \in \mathcal{F}$.

but

2) $A \subset B$ and $B \in \mathcal{G}$ together do *not* imply that $A \in \mathcal{G}$.

Example. Let $\Omega = [0,1]$ (cf. Example 1.3). Then $\mathcal{F}_3$ is a sub-σ-field of $\mathcal{B}([0,1])$.

1.2 Random variables

The notion of random variable is usually introduced in elementary probability courses as a vague concept, essentially characterized by its distribution. In mathematical terms however, random variables do exist prior to their distribution: they are functions from the fundamental set Ω to $\mathbb{R}$. Here is a preliminary definition.

Definition 1.6. On the set $\mathbb{R}$, one defines the *Borel σ-field* as

$$\mathcal{B}(\mathbb{R}) = \sigma(\{\,]a,b[\,:\ a,b \in \mathbb{R}, a < b\}).$$

The elements of $\mathcal{B}(\mathbb{R})$ are called *Borel sets.* Again, notice that $\mathcal{B}(\mathbb{R})$ is strictly included in $\mathcal{P}(\mathbb{R})$.

Definition 1.7. Let $(\Omega, \mathcal{F})$ be a measurable space. A *random variable* on $(\Omega, \mathcal{F})$ is a map $X : \Omega \to \mathbb{R}$ satisfying

$$\{\omega \in \Omega \,:\, X(\omega) \in B\} \in \mathcal{F}, \quad \forall B \in \mathcal{B}(\mathbb{R}). \tag{1}$$

Notation. One often simply denotes the set $\{\omega \in \Omega \,:\, X(\omega) \in B\} = \{X \in B\} = X^{-1}(B)$: it is called the inverse image of the set B through the map X (watch out that X need not be a bijective function in order for this set to be well defined).

Terminology. The above random variable X is sometimes called *$\mathcal{F}$-measurable*, in order to emphasize that if one knows the information $\mathcal{F}$, then one knows the value of X.

Example. If $\mathcal{F} = \mathcal{P}(\Omega)$, then condition (1) is always satisfied, so every map $X : \Omega \to \mathbb{R}$ is an $\mathcal{F}$-measurable random variable. On the contrary, if $\mathcal{F} = \{\emptyset, \Omega\}$, then the only random variables which are $\mathcal{F}$-measurable are the maps $X : \Omega \to \mathbb{R}$ which are constant.

Remark. Condition (1) can be shown to be equivalent to the following condition:

$$\{\omega \in \Omega \,:\, X(\omega) \leq t\} \in \mathcal{F}, \quad \forall t \in \mathbb{R},$$

which is significantly easier to check.

Definition 1.8. Let $(\Omega, \mathcal{F})$ be a measurable space and $A \in \mathcal{F}$ be an event. Then the map $\Omega \to \mathbb{R}$ defined as

$$\omega \mapsto 1_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{otherwise,} \end{cases}$$

is a random variable on $(\Omega, \mathcal{F})$. It is called the *indicator function* of the event A.

Example. Let $\Omega = \{1, \ldots, 6\}$ and $\mathcal{F} = \mathcal{P}(\Omega)$ (cf. Example 1.2). Then $X_1(\omega) = \omega$ and $X_2(\omega) = 1_{\{1,3,5\}}(\omega)$ are both random variables on $(\Omega, \mathcal{F})$. Moreover, X_2 is $\mathcal{F}_2$-measurable, but notice that X_1 is neither $\mathcal{F}_1$- nor $\mathcal{F}_2$-measurable.

Example. Let $\Omega = [0, 1]$ and $\mathcal{F} = \mathcal{B}([0, 1])$ (cf. Example 1.3). Then $X_3(\omega) = \sum_{j=1}^n x_j 1_{I_j}(\omega)$ and $X_4(\omega) = \omega$ are both random variables on $(\Omega, \mathcal{F})$. Notice however that only X_3 is $\mathcal{F}_3$-measurable.

We will need to consider not only random variables, but also functions of random variables. This is why we introduce the following definition.

Definition 1.9. A map $g : \mathbb{R} \to \mathbb{R}$ such that

$$\{x \in \mathbb{R} \,:\, g(x) \in B\} \in \mathcal{B}(\mathbb{R}), \quad \forall B \in \mathcal{B}(\mathbb{R}),$$

is called a *Borel-measurable function* on $\mathbb{R}$.

Remark. A Borel-measurable function on $\mathbb{R}$ is therefore nothing but a random variable on the measurable space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

As it is difficult to construct explicitly sets which are not Borel sets, it is equally difficult to construct functions which are not Borel-measurable. Nevertheless, one often needs to check that a given function is Borel-measurable. A useful criterion for this is the following (given here without proof).

Proposition 1.10. If $g : \mathbb{R} \to \mathbb{R}$ is continuous, then it is Borel-measurable.

Finally, let us mention this useful property of functions of random variables.

Proposition 1.11. If X is an $\mathcal{F}$-measurable random variable and $g : \mathbb{R} \to \mathbb{R}$ is Borel-measurable, then $Y = g(X)$ is also an $\mathcal{F}$-measurable random variable.

Proof. Let $B \in \mathcal{B}(\mathbb{R})$. Then

$$\{Y \in B\} = \{g(X) \in B\} = \{X \in g^{-1}(B)\} \in \mathcal{F},$$

since X is an $\mathcal{F}$-measurable random variable and $g^{-1}(B) \in \mathcal{B}(\mathbb{R})$ by assumption. □

σ-field generated by a collection of random variables.

The amount of information contained in a random variable, or more generally in a collection of random variables, is given by the definition below.

Definition 1.12. Let $(\Omega, \mathcal{F})$ be a measurable space and $\{X_i,\, i \in I\}$ be a collection of random variables on $(\Omega, \mathcal{F})$. The *σ-field generated by* X_i, $i \in I$, denoted as $\sigma(X_i,\, i \in I)$, is the smallest σ-field $\mathcal{G}$ on Ω such that all the random variables X_i are $\mathcal{G}$-measurable.

Remark. Notice that

$$\sigma(X_i,\, i \in I) = \sigma(\{\{X_i \in B\},\, i \in I,\, B \in \mathcal{B}(\mathbb{R})\}),$$

where the right-hand side expression refers to Definition 1.4. It turns out that one also has

$$\sigma(X_i,\, i \in I) = \sigma(\{\{X_i \le t\},\, i \in I,\, t \in \mathbb{R}\}).$$

Example. Let $(\Omega, \mathcal{F})$ be a measurable space. If X_0 is a constant random variable (i.e. $X_0(\omega) = c \in \mathbb{R},\ \forall \omega \in \Omega$), then $\sigma(X_0) = \{\emptyset, \Omega\}$.

Example. Let $\Omega = \{1, \dots, 6\}$ and $\mathcal{F} = \mathcal{P}(\Omega)$ (cf. Example 1.2). Then $\sigma(X_1) = \mathcal{P}(\Omega)$ and $\sigma(X_2) = \mathcal{F}_2$.

Example. Let $\Omega = [0, 1]$ and $\mathcal{F} = \mathcal{B}([0, 1])$ (cf. Example 1.3). Then $\sigma(X_3) = \mathcal{F}_3$ and $\sigma(X_4) = \mathcal{B}([0, 1])$.

Following the proof of Proposition 1.11, the proposition below can be easily shown.

Proposition 1.13. If X is a random variable on a measurable space $(\Omega, \mathcal{F})$ and $g : \mathbb{R} \to \mathbb{R}$ is Borel-measurable, then $Y = g(X)$ is a $\sigma(X)$-measurable random variable (this applies in particular to $Y = X$).

As a matter of fact, it turns out that the reciprocal statement is also true: if Y is a $\sigma(X)$-measurable random variable, then there exists a Borel-measurable function $g : \mathbb{R} \to \mathbb{R}$ such that $Y = g(X)$.

1.3 Probability measures

Definition 1.14. Let $(\Omega, \mathcal{F})$ be a measurable space. A *probability measure* on $(\Omega, \mathcal{F})$ is a map $\mathbb{P} : \mathcal{F} \to [0, 1]$ satisfying the following two axioms:

(i) $\mathbb{P}(\emptyset) = 0$ and $\mathbb{P}(\Omega) = 1$.

(ii) If $(A_n)_{n=1}^\infty \subset \mathcal{F}$ is such that $A_n \cap A_m = \emptyset$, $\forall n \ne m$, then $\mathbb{P}(\cup_{n=1}^\infty A_n) = \sum_{n=1}^\infty \mathbb{P}(A_n)$.
In particular, if $A, B \in \mathcal{F}$ are such that $A \cap B = \emptyset$, then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

The following properties can be further deduced from the above axioms:

(iii) If $(A_n)_{n=1}^\infty \subset \mathcal{F}$, then $\mathbb{P}(\cup_{n=1}^\infty A_n) \le \sum_{n=1}^\infty \mathbb{P}(A_n)$.
In particular, if $A, B \in \mathcal{F}$, then $\mathbb{P}(A \cup B) \le \mathbb{P}(A) + \mathbb{P}(B)$.

(iv) If $A, B \in \mathcal{F}$ and $A \subset B$, then $\mathbb{P}(A) \le \mathbb{P}(B)$ and $\mathbb{P}(B \backslash A) = \mathbb{P}(B) - \mathbb{P}(A)$.
In particular, $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.

(v) If $A, B \in \mathcal{F}$, then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

(vi) If $(A_n)_{n=1}^\infty \subset \mathcal{F}$ is such that $A_n \subset A_{n+1}$, $\forall n$, then $\mathbb{P}(\cup_{n=1}^\infty A_n) = \lim_{n\to\infty} \mathbb{P}(A_n)$.

(vii) If $(A_n)_{n=1}^\infty \subset \mathcal{F}$ is such that $A_n \supset A_{n+1}$, $\forall n$, then $\mathbb{P}(\cap_{n=1}^\infty A_n) = \lim_{n\to\infty} \mathbb{P}(A_n)$.

Terminology. The triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a *probability space*. Property (ii) is referred to as the *σ-additivity* (or simply *additivity* in the finite case) of probability measures.

Example. Let $\Omega = \{1, .., 6\}$ and $\mathcal{F} = \mathcal{P}(\Omega)$ be the measurable space associated to a die roll. The probability measure associated to a balanced die is defined as

$$\mathbb{P}_1(\{i\}) = \frac{1}{6},\ \forall i \in \{1, \dots, 6\},$$

and is extended by additivity to all subsets of Ω. E.g.,

$$\mathbb{P}_1(\{1,3,5\}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}.$$

The probability measure associated to a loaded die is defined as

$$\mathbb{P}_2(\{6\}) = 1 \quad \text{and} \quad \mathbb{P}_2(\{i\}) = 0,\ \forall i \in \{1,\dots,5\},$$

and is extended by additivity to all subsets of Ω.

Example. Let $\Omega = [0,1]$ and $\mathcal{F} = \mathcal{B}([0,1])$. One defines the following probability measure on the subintervals of $[0,1]$:

$$\mathbb{P}(\,]a,b[\,) = b - a.$$

Fact. $\mathbb{P}$ can be extended by σ-additivity to all Borel subsets of $[0,1]$. It is called the *Lebesgue measure* on $[0,1]$ and is sometimes denoted as $\mathbb{P}(B) = |B|$.

1.4 Distribution of a random variable

Definition 1.15. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X be a random variable defined on this probability space. The *distribution* of X is the map $\mu_X : \mathcal{B}(\mathbb{R}) \to [0,1]$ defined as

$$\mu_X(B) = \mathbb{P}(\{X \in B\}), \quad B \in \mathcal{B}(\mathbb{R}).$$

Remark. The triple $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu_X)$ forms a new probability space.

Notation. If a random variable X has distribution μ, this is denoted as $X \sim \mu$. Likewise, if two random variables X and Y share the same distribution μ, then they are are said to be *identically distributed* and this is denoted as $X \sim Y \sim \mu$.

Example 1.16. The probability space describing two independent (and balanced) dice rolls is $\Omega = \{1,\dots,6\} \times \{1,\dots,6\}$, $\mathcal{F} = \mathcal{P}(\Omega)$ and

$$\mathbb{P}(\{(i,j)\}) = \frac{1}{36}, \quad \forall (i,j) \in \Omega.$$

Let $X_1(i,j) = i$ be the result of the first die, and $Y(i,j) = i + j$ be the sum of the two dice. Then

$$\mu_{X_1}(\{i\}) = \mathbb{P}(\{X_1 = i\}) = \mathbb{P}(\{(i,1),\dots,(i,6)\}) = \frac{6}{36} = \frac{1}{6}, \quad \forall i \in \{1,\dots,6\},$$

and

$$\mu_Y(\{2\}) = \mathbb{P}(\{Y = 2\}) = \mathbb{P}(\{(1,1)\}) = \frac{1}{36}, \quad \mu_Y(\{3\}) = \mathbb{P}(\{Y = 3\}) = \mathbb{P}(\{(1,2),(2,1)\}) = \frac{1}{18}.$$

More generally:

$$\mu_Y(\{i\}) = \frac{6 - |7 - i|}{36}, \quad i \in \{2,\dots,12\}.$$

Cumulative distribution function.

Definition 1.17. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X be a random variable defined on this probability space. The *cumulative distribution function (or cdf)* of X is the map $F_X : \mathbb{R} \to [0,1]$ defined as

$$F_X(t) = \mu_X(\,]-\infty, t]) = \mathbb{P}(\{X \le t\}), \quad t \in \mathbb{R}.$$

Fact. The knowledge of F_X is equivalent to the knowledge of μ_X.

From the properties of probability measures, one deduces easily that the cdf of a random variable satisfies the following properties:

(i) $\lim_{t\to-\infty} F_X(t) = 0$, $\lim_{t\to+\infty} F_X(t) = 1$.

(ii) F_X is *non-decreasing*, i.e. $F_X(s) \le F_X(t)$ for all $s < t$.

(iii) F_X is *right-continuous* on $\mathbb{R}$, i.e. $\lim_{\varepsilon\downarrow 0} F_X(t+\varepsilon) = F_X(t)$, for all $t \in \mathbb{R}$.

Remark. F_X has at most a countable number of jumps on the real line. If F_X has a jump of size $p \in [0,1]$ at $t \in \mathbb{R}$, this actually means that $\mathbb{P}(\{X = t\}) = F_X(t) - \lim_{\varepsilon\downarrow 0} F_X(t-\varepsilon) = p$.

Two important classes of random variables.

Discrete random variables.

Definition 1.18. X is a *discrete random variable* if it takes values in a countable subset C of $\mathbb{R}$, that is, $\mathbb{P}(\{X \in C\}) = 1$.

The distribution of a discrete random variable is entirely characterized by the numbers $p_x = \mathbb{P}(\{X = x\})$, where $x \in C$. Notice that $p_x \ge 0$ for all $x \in C$ and that $\sum_{x\in C} p_x = \mathbb{P}(\{X \in C\}) = 1$. Moreover,

$$\mu_X(B) = \mathbb{P}(\{X \in B\}) = \sum_{x\in B} p_x, \quad \forall B \in \mathcal{B}(\mathbb{R}),$$

and

$$F_X(t) = \mathbb{P}(\{X \le t\}) = \sum_{x\le t} p_x, \quad \forall t \in \mathbb{R},$$

is a step function.

Example. A binomial random variable X with parameters $n \ge 1$ and $p \in [0,1]$ (denoted as $X \sim \mathrm{Bi}(n,p)$) takes values in $\{0,\ldots,n\}$ and is characterized by the numbers

$$p_k = \mathbb{P}(\{X = k\}) = \begin{pmatrix} n \\ k \end{pmatrix} p^k\,(1-p)^{n-k}, \quad k \in \{0,\ldots,n\},$$

where $\begin{pmatrix} n \\ k \end{pmatrix} = \dfrac{n!}{k!(n-k)!}$ are the binomial coefficients.

Continuous random variables.

Definition 1.19. X is a *continuous random variable* if $\mathbb{P}(\{X \in B\}) = 0$ whenever $B \in \mathcal{B}(\mathbb{R})$ is such that $|B| = 0$ (remember that $|B|$ is the Lebesgue measure of B).

In particular, this implies that if X is a continuous random variable, then $\mathbb{P}(\{X = x\}) = 0$ $\forall x \in \mathbb{R}$ (as $|\{x\}| = 0$ $\forall x \in \mathbb{R}$).

Fact. If X is a continuous random variable according to the above definition, then there exists a function $f_X : \mathbb{R} \to \mathbb{R}$, called the *probability density function (or pdf)* of X, such that $f_X(x) \ge 0$ $\forall x \in \mathbb{R}$, $\int_{\mathbb{R}} f_X(x)\,dx = 1$ and

$$\mu_X(B) = \mathbb{P}(\{X \in B\}) = \int_B f_X(x)\,dx, \quad \forall B \in \mathcal{B}(\mathbb{R}).$$

Moreover,

$$F_X(t) = \mathbb{P}(\{X \le t\}) = \int_{-\infty}^t f_X(x)\,dx, \quad \forall t \in \mathbb{R},$$

is a differentiable function (whose derivative is $F_X'(t) = f_X(t)$).

Example. A Gaussian random variable X with mean μ and variance σ^2 (denoted as $X \sim \mathcal{N}(\mu, \sigma^2)$) takes values in $\mathbb{R}$ and has pdf

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}.$$

1.5 Independence

The notion of independence is a central notion in probability. It is usually defined for events and random variables in elementary probability courses. Nevertheless, as it will become clear below, the independence between *σ-fields* turns out to be the most natural concept (remembering that a σ-field is related to the amount of information one has on a system).

In the three paragraphs below, $(\Omega, \mathcal{F}, \mathbb{P})$ denotes a generic probability space.

Independence of events.

One starts by defining the independence of two events in $\mathcal{F}$.

Definition 1.20. Two events $A, B \in \mathcal{F}$ are independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$.

Notation. $A \perp\!\!\!\perp B$.

Proposition 1.21. If two events $A, B \in \mathcal{F}$ are independent, then it also holds that

$$\mathbb{P}(A \cap B^c) = \mathbb{P}(A)\,\mathbb{P}(B^c), \quad \mathbb{P}(A^c \cap B) = \mathbb{P}(A^c)\,\mathbb{P}(B) \quad \text{and} \quad \mathbb{P}(A^c \cap B^c) = \mathbb{P}(A^c)\,\mathbb{P}(B^c).$$

Proof. One shows here the first equality (noticing that the other two can be proved in a similar way):

$$\mathbb{P}(A \cap B^c) = \mathbb{P}(A \backslash (A \cap B)) = \mathbb{P}(A) - \mathbb{P}(A \cap B) = \mathbb{P}(A) - \mathbb{P}(A)\,\mathbb{P}(B) = \mathbb{P}(A)\,(1 - \mathbb{P}(B)) = \mathbb{P}(A)\,\mathbb{P}(B^c).$$

□

For a collection of more than 2 events, the property $\mathbb{P}(A_1 \cap \ldots \cap A_n) = \mathbb{P}(A_1) \cdots \mathbb{P}(A_n)$ does not suffice to guarantee that the same property holds for complements of the events A_i. A slightly more involved definition of independence is therefore required.

Definition 1.22. Let $\{A_1, \ldots, A_n\}$ be a collection of events in $\mathcal{F}$. This collection is independent if

$$\mathbb{P}(A_1^* \cap \ldots \cap A_n^*) = \mathbb{P}(A_1^*) \cdots \mathbb{P}(A_n^*)$$

where $A_i^* =$ either A_i or A_i^c, $i \in \{1, \ldots, n\}$.

An intuitive reason why complements should be included in the definition of independence is the following. Let us assume that one rolls a balanced die with four faces. Then the events {the outcome is 1 or 2} and {the outcome is even} are clearly independent; more precisely, the different *informations* associated with these events are. So the events {the outcome is 1 or 2} and {the outcome is odd} are also independent. This motivates the extension of the definition of independence to σ-fields in the next paragraph.

Fact. It can be shown that Definition 1.22 is equivalent to saying that

$$\mathbb{P}\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} \mathbb{P}(A_i), \quad \forall I \subset \{1, \ldots, n\}.$$

From the above fact, one deduces that a collection of events might not be independent, even though its events are two-by-two independent.

Independence of σ-fields.

Definition 1.23. Let $\{\mathcal{G}_1, \ldots, \mathcal{G}_n\}$ be a collection of sub-σ-fields of $\mathcal{F}$. This collection is independent if

$$\mathbb{P}(A_1 \cap \ldots \cap A_n) = \mathbb{P}(A_1) \cdots \mathbb{P}(A_n), \quad \forall A_1 \in \mathcal{G}_1, \ldots, A_n \in \mathcal{G}_n.$$

Example. Let again $\{A_1, \ldots, A_n\}$ be a collection of events in $\mathcal{F}$. Then the collection of events $\{A_1, \ldots, A_n\}$ is independent (according to Definition 1.22) if and only if the collection of σ-fields $\{\sigma(A_1), \ldots, \sigma(A_n)\}$ is independent (according to Definition 1.23). In order to see this, observe that $\sigma(A_i) = \{\emptyset, A_i, A_i^c, \Omega\}$.

Independence of random variables.

Definition 1.24. Let $\{X_1, \ldots, X_n\}$ be a collection of random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$. This collection is independent if the collection of σ-fields $\{\sigma(X_1), \ldots, \sigma(X_n)\}$ is independent.

Since $\sigma(X_i) = \sigma(\{X_i \in B\}, B \in \mathcal{B}(\mathbb{R}))$, the collection $\{X_1, \ldots, X_n\}$ is independent if and only if

$$\mathbb{P}(\{X_1 \in B_1, \ldots, X_n \in B_n\}) = \mathbb{P}(\{X_1 \in B_1\}) \cdots \mathbb{P}(\{X_n \in B_n\}), \quad \forall B_1, \ldots, B_n \in \mathcal{B}(\mathbb{R}).$$

But one also knows that $\sigma(X_i) = \sigma(\{X_i \leq t\}, t \in \mathbb{R})$, so it turns out that $\{X_1, \ldots, X_n\}$ is independent if and only if

$$\mathbb{P}(\{X_1 \leq t_1, \ldots, X_n \leq t_n\}) = \mathbb{P}(\{X_1 \leq t_1\}) \cdots \mathbb{P}(\{X_n \leq t_n\}), \quad \forall t_1, \ldots, t_n \in \mathbb{R}.$$

For discrete random variables taking values in a countable set C, this reduces to

$$\mathbb{P}(\{X_1 = x_1, \ldots, X_n = x_n\}) = \mathbb{P}(\{X_1 = x_1\}) \cdots \mathbb{P}(\{X_n = x_n\}), \quad \forall x_1, \ldots, x_n \in C.$$

And for jointly continuous random variables with joint pdf $f_{X_1,\ldots,X_n}$, this reduces to the classical relation

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n), \quad \forall x_1, \ldots, x_n \in \mathbb{R}.$$

The advantage of the above theoretical definition involving σ-fields is the following. Assume $\{X_1, \ldots, X_n\}$ is a collection of independent random variables and let $g_1, \ldots, g_n : \mathbb{R} \to \mathbb{R}$ be Borel-measurable functions. Then one directly deduces from the definition (and the fact that $g_i(X_i)$ is $\sigma(X_i)$-measurable) that $\{g_1(X_1), \ldots, g_n(X_n)\}$ is also a collection of independent random variables, which might have been cumbersome to check using any of the other "simpler" definition.

Example. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a generic probability space and let $X_0(\omega) = c \in \mathbb{R}$, $\forall \omega \in \Omega$ be a constant random variable. As $\sigma(X_0) = \mathcal{F}_0 = \{\emptyset, \Omega\}$, X_0 is independent of any other random variable defined on $(\Omega, \mathcal{F}, \mathbb{P})$.

Example. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the probability space describing two independent dice rolls in Example 1.16 and let $X_1(i,j) = i$ and $X_2(i,j) = j$. One verifies below that these two random variables are indeed independent. It was already shown that $\mathbb{P}(\{X_1 = i\}) = \frac{1}{6}$, $\forall i \in \{1, \ldots, 6\}$. Likewise, $\mathbb{P}(\{X_2 = j\}) = \frac{1}{6}$, $\forall j \in \{1, \ldots, 6\}$ and

$$\mathbb{P}(\{X_1 = i, X_2 = j\}) = \mathbb{P}(\{(i,j)\}) = \frac{1}{36} = \mathbb{P}(\{X_1 = i\})\, \mathbb{P}(\{X_2 = j\}), \quad \forall (i,j) \in \Omega,$$

so X_1 and X_2 are independent.

1.6 Expectation

From the point of view of measure theory, random variables are maps from Ω to $\mathbb{R}$. Correspondingly, the *expectation* (or *mean*) of a random variable X is the *Lebesgue integral* of the map X, that is, the "area under the curve $\omega \mapsto X(\omega)$", where the horizontal axis is measured with the probability measure $\mathbb{P}$.

Definition.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X be a random variable defined on this probability space. The expectation of X, denoted as $\mathbb{E}(X)$, will be defined in three steps.

Step 1. Assume first that X is a non-negative discrete random variable, i.e. that X may be written as

$$X(\omega) = \sum_{i=i}^{\infty} x_i \, 1_{A_i}(\omega),$$

where $x_i \geq 0$ and $A_i \in \mathcal{F}$ (notice that if the x_i are all different, then $A_i = \{X = x_i\}$). The expectation of X is then defined as

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

which corresponds to the traditional definition of expectation in elementary probability courses. Notice here that since the sum is infinite, $\mathbb{E}(X)$ may take the value $+\infty$; but because of the assumption that $x_i \geq 0$, $\mathbb{E}(X)$ is always non-negative.

Notice also that in the particular case where $X = 1_A$, with $A \in \mathcal{F}$, one has $\mathbb{E}(X) = \mathbb{P}(A)$.

Step 2. Assume now that X is a generic non-negative random variable (i.e. $X(\omega) \geq 0$, $\forall \omega \in \Omega$). Let us define the following sequence of discrete random variables:

$$X_n(\omega) = \sum_{i=1}^{\infty} \frac{i-1}{2^n} \, 1_{\{\frac{i-1}{2^n} < X \leq \frac{i}{2^n}\}}(\omega).$$

Notice that $x_i = \frac{i-1}{2^n} \geq 0$ and that $\{\frac{i-1}{2^n} < X \leq \frac{i}{2^n}\} \in \mathcal{F}$, since X is $\mathcal{F}$-measurable. So according to Step 1, one has for each n

$$\mathbb{E}(X_n) = \sum_{i=1}^{\infty} \frac{i-1}{2^n} \, \mathbb{P}\left(\left\{\frac{i-1}{2^n} < X \leq \frac{i}{2^n}\right\}\right) \in [0, +\infty].$$

It should be observed that $(X_n, \, n \in \mathbb{N})$ is actually an increasing sequence of non-negative "staircases", that is,

$$0 \leq X_n(\omega) \leq X_{n+1}(\omega), \quad \forall n.$$

As the size of the steps is divided by two from n to $n+1$, the staircase gets refined. Likewise, one easily sees that $\mathbb{E}(X_n) \leq \mathbb{E}(X_{n+1})$ for all n, so $(\mathbb{E}(X_n), n \in \mathbb{N})$ is an increasing sequence, that therefore converges (possibly to $+\infty$). One defines

$$\mathbb{E}(X) = \lim_{n \to \infty} \mathbb{E}(X_n) = \lim_{n \to \infty} \sum_{i=1}^{\infty} \frac{i-1}{2^n} \, \mathbb{P}\left(\left\{\frac{i-1}{2^n} < X \leq \frac{i}{2^n}\right\}\right) \in [0, \infty].$$

Step 3. Finally, consider a generic random variable X. One defines its *positive and negative parts*:

$$X^+(\omega) = \max(0, X(\omega)), \quad X^-(\omega) = \max(0, -X(\omega))$$

Notice that both $X^+(\omega) \geq 0$ and $X^-(\omega) \geq 0$, and that

$$X^+(\omega) - X^-(\omega) = X(\omega), \quad X^+(\omega) + X^-(\omega) = |X(\omega)|.$$

In measure theory, one does not want to deal with ill defined quantities such as $\infty - \infty$. One therefore defines $\mathbb{E}(X)$ only when $\mathbb{E}(|X|) = \mathbb{E}(X^+) + \mathbb{E}(X^-) < \infty$:

$$\mathbb{E}(X) = \mathbb{E}(X^+) - \mathbb{E}(X^-).$$

Two important particular cases. Let X be a random variable and $g : \mathbb{R} \to \mathbb{R}$ be a Borel-measurable function such that $\mathbb{E}(|g(X)|) < \infty$ (this last condition is verified if for example g is a bounded function).

- If X is a discrete random variable with values in a countable set C, then

$$\mathbb{E}(g(X)) = \sum_{x \in C} g(x)\, \mathbb{P}(\{X = x\}).$$

- If X is a continuous random variable with pdf f_X, then

$$\mathbb{E}(g(X)) = \int_{\mathbb{R}} g(x)\, f_X(x)\, dx.$$

Terminology. - If $\mathbb{E}(|X|) < \infty$, then X is said to be an *integrable* random variable.
- If $\mathbb{E}(X^2) < \infty$, then X is said to be a *square-integrable* random variable.
- If there exists $c > 0$ such that $|X(\omega)| \le c$, $\forall \omega \in \Omega$, then X is said to be a *bounded* random variable.
- If $\mathbb{E}(X) = 0$, then X is said to be a *centered* random variable.

One has the following series of implications:

$$\begin{gathered} X \text{ is bounded} \quad \Rightarrow \quad X \text{ is square-integrable} \quad \Rightarrow \quad X \text{ is integrable}, \\ X \text{ is integrable and } Y \text{ is bounded} \quad \Rightarrow \quad XY \text{ is integrable}, \\ X, Y \text{ are both square-integrable} \quad \Rightarrow \quad XY \text{ is integrable.} \end{gathered}$$

Negligible and almost sure sets. An event $A \in \mathcal{F}$ is said to be negligible if $\mathbb{P}(A) = 0$. On the contrary, an event $B \in \mathcal{F}$ is said to be almost sure (a.s.) if $\mathbb{P}(B) = 1$. For example, if $\mathbb{P}(\{X \ge c\}) = 1$, one says that "$X \ge c$ almost surely".

Basic properties of expectation.

Linearity. If $c \in \mathbb{R}$ and X, Y are integrable, then $\mathbb{E}(cX + Y) = c\,\mathbb{E}(X) + \mathbb{E}(Y)$.

Positivity. If X is integrable and $X \ge 0$ a.s., then $\mathbb{E}(X) \ge 0$.

Strict positivity. If X is integrable, $X \ge 0$ a.s. and $\mathbb{E}(X) = 0$, then $X = 0$ a.s.

Monotonicity. If X, Y are integrable and $X \ge Y$ a.s., then $\mathbb{E}(X) \ge \mathbb{E}(Y)$.

Inequalities.

Cauchy-Schwarz's inequality. If X, Y are square-integrable random variables, then the product XY is integrable and

$$\mathbb{E}(|XY|) \le \sqrt{\mathbb{E}(X^2)}\, \sqrt{\mathbb{E}(Y^2)}.$$

In particular, considering $Y = 1$ shows that if X is square-integrable, then it is also integrable.

Jensen's inequality. If X is a random variable and $\psi : \mathbb{R} \to \mathbb{R}$ is convex and such that $\mathbb{E}(|\psi(X)|) < \infty$, then

$$\psi(\mathbb{E}(X)) \le \mathbb{E}(\psi(X)).$$

In particular, $|\mathbb{E}(X)| \le \mathbb{E}(|X|)$.

Also, if X is such that $\mathbb{P}(\{X = a\}) = \mathbb{P}(\{X = b\}) = 1/2$, then the above inequality says that

$$\psi\left(\frac{a+b}{2}\right) \le \frac{\psi(a) + \psi(b)}{2},$$

which is pretty much the definition of convexity for ψ.

Chebychev's inequality. If X is a random variable and $\varphi : \mathbb{R} \to \mathbb{R}_+$ is increasing on $\mathbb{R}_+$ and such that $\mathbb{E}(\varphi(X)) < \infty$, then for any $a > 0$, one has

$$\mathbb{P}(\{X \geq a\}) \leq \frac{\mathbb{E}(\varphi(X))}{\varphi(a)}.$$

In particular, if X is square-integrable, then taking $\varphi(x) = x^2$ gives

$$\mathbb{P}(\{X \geq a\}) \leq \frac{\mathbb{E}(X^2)}{a^2}.$$

Variance, covariance and independence.

Definition 1.25. Let X, Y be two square-integrable random variables. The *variance* of X is defined as

$$\mathrm{Var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 \geq 0$$

and the *covariance* of X and Y is defined as

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

Terminology. If $\mathrm{Cov}(X, Y) = 0$, then X and Y are said to be *uncorrelated.*

Fact. If X, Y are independent square-integrable random variables, then

a) $\mathrm{Cov}(X, Y) = 0$, i.e. X and Y are uncorrelated (but the reciprocal statement is wrong).

b) $\mathrm{Var}(cX + Y) = c^2\,\mathrm{Var}(X) + \mathrm{Var}(Y)$, for any $c \in \mathbb{R}$.

1.7 Convergence of sequences of random variables

For a given sequence of random variables $(X_n,\, n \geq 1)$ defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$, there are several notions of convergence to a limiting random variable X. Let us review the most important ones.

Convergence in probability. The sequence (X_n) is said to converge in probability to X (and this is denoted as $X_n \xrightarrow{\mathbb{P}} X$) if for all $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(\{|X_n - X| > \varepsilon\}) = 0.$$

Almost sure convergence. The sequence (X_n) is said to converge almost surely to X (and this is denoted as $X_n \to X$ a.s.) if

$$\mathbb{P}\left(\left\{\lim_{n\to\infty} X_n = X\right\}\right) = 1.$$

Fact. Almost sure convergence implies convergence in probability, but the reverse implication is wrong. Nevertheless, it holds that $X_n \to X$ a.s. if for all $\varepsilon > 0$,

$$\sum_{n=1}^{\infty} \mathbb{P}(\{|X_n - X| > \varepsilon\}) < \infty.$$

Quadratic convergence. Let us moreover assume that the random variables X_n and X are square-integrable. The sequence (X_n) is then said to converge quadratically to X if

$$\lim_{n\to\infty} \mathbb{E}(|X_n - X|^2) = 0.$$

Fact. By Chebychev's inequality, quadratic convergence implies convergence in probability (but not almost sure convergence).

Convergence in distribution. The sequence (X_n) is said to converge in distribution to X (and this is denoted as $X_n \xrightarrow{d} X$) if

$$\lim_{n\to\infty} F_{X_n}(t) = F_X(t),$$

for all $t \in \mathbb{R}$ which are continuity points of F_X.

Remark. For this last definition, the random variables X_n need not be all defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The knowledge of their respective distributions suffices.

"Examples": limit theorems.

Weak law of large numbers (not the standard version). Let $(\xi_n,\, n \geq 1)$ be a sequence of square-integrable and uncorrelated random variables with a common expectation $\mathbb{E}(\xi_n) = \mu$ and a common variance $\mathrm{Var}(\xi_n) = \sigma^2$. Let also $S_n = \xi_1 + \ldots + \xi_n$. Then

$$\frac{S_n}{n} \xrightarrow{\mathbb{P}} \mu.$$

Remark. The convergence is also quadratic in this case.

Strong law of large numbers. Let $(\xi_n,\, n \geq 1)$ be a sequence independent and identically distributed (i.i.d.) random variables such that $\mathbb{E}(|\xi_1|) < \infty$. Let also $\mu = \mathbb{E}(\xi_1)$ and $S_n = \xi_1 + \ldots + \xi_n$. Then

$$\frac{S_n}{n} \to \mu \quad \text{a.s.}$$

Example. Assume that $\mathbb{P}(\{\xi_1 = 1\}) = \mathbb{P}(\{\xi_1 = 0\}) = 1/2$ (so $\mu = 1/2$). Then the above theorem says approximately that as n gets large,

$$S_n \simeq \frac{n}{2} \quad \text{with high probability.}$$

The next question is: for a given n, how close is S_n from $n/2$? The answer is given by the following theorem.

Central limit theorem. Let (ξ_n) be a sequence of i.i.d. random variables such that $\mathbb{E}(\xi_1^2) < \infty$. Let also $\mu = \mathbb{E}(\xi_1)$, $\sigma^2 = \mathrm{Var}(\xi_1)$ and $S_n = \xi_1 + \ldots + \xi_n$. Then

$$\frac{S_n - n\,\mu}{\sqrt{n}\,\sigma} \xrightarrow{d} Z \sim \mathcal{N}(0,1).$$

This more specifically says that

$$\lim_{n\to\infty} \mathbb{P}\left(\left\{\frac{S_n - n\,\mu}{\sqrt{n}\,\sigma} \leq t\right\}\right) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx,$$

for all $t \in \mathbb{R}$ (as the cdf of $\mathcal{N}(0,1)$ is continuous on $\mathbb{R}$).

Example. Assume again that $\mathbb{P}(\{\xi_1 = 1\}) = \mathbb{P}(\{\xi_1 = 0\}) = 1/2$ (so $\mu = 1/2$ and $\sigma = 1/2$). Then the above theorem says approximately that as n gets large,

$$S_n \simeq \frac{n}{2} + \frac{\sqrt{n}}{2}\, Z,$$

where Z is a standard Gaussian random variable. So typically, the standard deviation of S_n from its mean $n/2$ is of order $\sqrt{n}$.

1.8 Conditional expectation

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, as usual.

Conditioning with respect to an event $B \in \mathcal{F}$.

The conditional probability of an event $A \in \mathcal{F}$ given another event $B \in \mathcal{F}$ is defined as

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \quad \text{given that } \mathbb{P}(B) > 0.$$

In a similar way, the conditional expectation of an integrable random variable X given B is defined as

$$\mathbb{E}(X|B) = \frac{\mathbb{E}(X\, 1_B)}{\mathbb{P}(B)}, \quad \text{given that } \mathbb{P}(B) > 0.$$

Conditioning with respect to a discrete random variable Y.

Let us assume that the random variable Y (is $\mathcal{F}$-measurable and) takes values in a countable set C.

$$\begin{aligned} \mathbb{P}(A|Y) &= \varphi(Y), \qquad \text{where } \varphi(y) = \mathbb{P}(A|\{Y = y\}), \quad y \in C. \\ \mathbb{E}(X|Y) &= \psi(Y), \qquad \text{where } \psi(y) = \mathbb{E}(X|\{Y = y\}), \quad y \in C. \end{aligned}$$

If X is also a discrete random variable with values in C, then

$$\mathbb{E}(X|Y) = \psi(Y), \text{ where } \psi(y) = \frac{\mathbb{E}(X\, 1_{\{Y=y\}})}{\mathbb{P}(\{Y = y\})} = \sum_{x \in C} x\, \frac{\mathbb{E}(1_{\{X=x\}\cap\{Y=y\}})}{\mathbb{P}(\{Y = y\})} = \sum_{x \in C} x\, \mathbb{P}(\{X = x\}|\{Y = y\}).$$

Important remark. $\varphi(y)$ and $\psi(y)$ are regular functions, but $\mathbb{P}(A|Y)$ and $\mathbb{E}(X|Y)$ are *random variables*. They both are functions of the outcome of the random variable Y, that is, they are $\sigma(Y)$-measurable random variables.

Example. Let X_1, X_2 be two independent dice rolls and let us compute $\mathbb{E}(X_1 + X_2|X_2) = \psi(X_2)$, where

$$\begin{aligned} \psi(y) &= \mathbb{E}(X_1 + X_2|\{X_2 = y\}) = \frac{\mathbb{E}((X_1 + X_2)\, 1_{\{X_2=y\}})}{\mathbb{P}(\{X_2 = y\})} \\ &= \frac{\mathbb{E}(X_1\, 1_{\{X_2=y\}}) + \mathbb{E}(X_2\, 1_{\{X_2=y\}})}{\mathbb{P}(\{X_2 = y\})} \overset{(a)}{=} \frac{\mathbb{E}(X_1)\, \mathbb{E}(1_{\{X_2=y\}}) + \mathbb{E}(y\, 1_{\{X_2=y\}})}{\mathbb{P}(\{X_2 = y\})} \\ &= \frac{\mathbb{E}(X_1)\, \mathbb{P}(\{X_2 = y\}) + y\, \mathbb{P}(\{X_2 = y\})}{\mathbb{P}(\{X_2 = y\})} = \mathbb{E}(X_1) + y, \end{aligned}$$

where the independence assumption between X_1 and X_2 has been used in equality (a). So finally (as one would expect), $\mathbb{E}(X_1 + X_2|X_2) = \mathbb{E}(X_1) + X_2$, which can be explained intuitively as follows: the expectation of X_1 conditioned on X_2 is nothing but the expectation of X_1, as the outcome of X_2 provides no information on the outcome of X_1 (X_1 and X_2 being independent); on the other side, the expectation of X_2 conditioned on X_2 is exactly X_2, as the outcome of X_2 is known.

Conditioning with respect to a continuous random variable Y?

In this case, one faces the following problem: if Y is a continuous random variable, $\mathbb{P}(\{Y = y\}) = 0$ for all $y \in \mathbb{R}$. So a direct generalization of the above formulas to the continuous case is impossible at first sight. A possible solution to this problem is to replace the event $\{Y = y\}$ by $\{y \leq Y < y + \varepsilon\}$ and take the limit $\varepsilon \to 0$ for the definition of conditional expectation. This actually works, but also leads to a paradox in the multidimensional setting (known as Borel's paradox). In addition, some random variables are neither discrete, nor continuous. It turns out that the cleanest way to define conditional expectation in the general case is through σ-fields.

Conditioning with respect to a sub-σ-field $\mathcal{G}$.

In order to define the conditional expectation in the general case, one needs the following proposition.

Proposition 1.26. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$ and X be an integrable random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. There exists then an integrable random variable Z such that

(i) Z is $\mathcal{G}$-measurable,

(ii) $\mathbb{E}(ZU) = \mathbb{E}(XU)$ for any random variable U $\mathcal{G}$-measurable and bounded.

Moreover, if Z_1, Z_2 are two integrable random variables satisfying (i) and (ii), then $Z_1 = Z_2$ a.s.

Definition 1.27. The above random variable Z is called the *conditional expectation of X given $\mathcal{G}$*. It is defined up to a negligible set.

Notation. Z is denoted as $\mathbb{E}(X|\mathcal{G})$.

One further *defines* $\mathbb{P}(A|\mathcal{G}) = \mathbb{E}(1_A|\mathcal{G})$ for $A \in \mathcal{F}$.

Remark. Notice that as before, both $\mathbb{P}(A|\mathcal{G})$ and $\mathbb{E}(X|\mathcal{G})$ are random variables.

Properties.

The above definition does not give a computation rule for the conditional expectation; it is only an existence theorem. The properties listed below will therefore be of help for computing conditional expectations.

- Linearity. $\mathbb{E}(c\,X + Y|\mathcal{G}) = c\,\mathbb{E}(X|\mathcal{G}) + \mathbb{E}(Y|\mathcal{G})$ a.s.
- Monotonicity. If $X \geq Y$ a.s., then $\mathbb{E}(X|\mathcal{G}) \geq \mathbb{E}(Y|\mathcal{G})$ a.s.
- $\mathbb{E}(\mathbb{E}(X|\mathcal{G})) = \mathbb{E}(X)$.
- If X is independent of $\mathcal{G}$, then $\mathbb{E}(X|\mathcal{G}) = \mathbb{E}(X)$ a.s.
- If X is $\mathcal{G}$-measurable, then $\mathbb{E}(X|\mathcal{G}) = X$ a.s.
- If Y is $\mathcal{G}$-measurable and bounded, then $\mathbb{E}(XY|\mathcal{G}) = \mathbb{E}(X|\mathcal{G})\,Y$ a.s.
- If $\mathcal{H}$ is a sub-σ-field of $\mathcal{G}$, then $\mathbb{E}(\mathbb{E}(X|\mathcal{H})|\mathcal{G}) = \mathbb{E}(\mathbb{E}(X|\mathcal{G})|\mathcal{H}) = \mathbb{E}(X|\mathcal{H})$ a.s.

Some of these properties are illustrated below with an example.

Example. Let $\Omega = \{1, \ldots, 6\}$, $\mathcal{F} = \mathbb{P}(\Omega)$ and $\mathbb{P}(\{\omega\}) = \frac{1}{6}$ for $\omega = 1, \ldots, 6$ (the probability space of the die roll). Let also $X(\omega) = \omega$ be the outcome of the die roll and consider the two sub-σ-fields:

$$\mathcal{G} = \sigma(\{1,3\}, \{2\}, \{5\}, \{4,6\}) \quad \text{and} \quad \mathcal{H} = \sigma(\{1,3,5\}, \{2,4,6\}).$$

Then $\mathbb{E}(X) = 3.5$,

$$\mathbb{E}(X|\mathcal{G})(\omega) = \begin{cases} 2 & \text{if } \omega \in \{1,3\} \text{ or } \omega = 2 \\ 5 & \text{if } \omega \in \{4,6\} \text{ or } \omega = 5 \end{cases} \quad \text{and} \quad \mathbb{E}(X|\mathcal{H})(\omega) = \begin{cases} 3 & \text{if } \omega \in \{1,3,5\} \\ 4 & \text{if } \omega \in \{2,4,6\} \end{cases}$$

So $\mathbb{E}(\mathbb{E}(X|\mathcal{G})) = \mathbb{E}(\mathbb{E}(X|\mathcal{H})) = \mathbb{E}(X)$. Moreover,

$$\mathbb{E}(\mathbb{E}(X|\mathcal{G})|\mathcal{H})(\omega) = \begin{cases} \frac{1}{3}(2+2+5) = 3 & \text{if } \omega \in \{1,3,5\} \\ \frac{1}{3}(2+5+5) = 4 & \text{if } \omega \in \{2,4,6\} \end{cases} = \mathbb{E}(X|\mathcal{H})(\omega)$$

and

$$\mathbb{E}(\mathbb{E}(X|\mathcal{H})|\mathcal{G})(\omega) = \begin{cases} 3 & \text{if } \omega \in \{1,3\} \text{ or } \omega = 5 \\ 4 & \text{if } \omega \in \{4,6\} \text{ or } \omega = 2 \end{cases} = \mathbb{E}(X|\mathcal{H})(\omega).$$

On other words, the smallest σ-field always “wins”.

Proposition 1.28. Let $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$, X, Y be two random variables such that X is independent of $\mathcal{G}$ and Y is $\mathcal{G}$-measurable, an let $\varphi : \mathbb{R}^2 \to \mathbb{R}$ be a Borel-measurable function such that $\mathbb{E}(|\varphi(X,Y)|) < \infty$. Then

$$\mathbb{E}(\varphi(X,Y)|\mathcal{G}) = \psi(Y) \quad a.s., \quad \text{where } \psi(y) = \mathbb{E}(\varphi(X,y)).$$

This proposition has the following consequence: when computing the expectation of a function φ of two independent random variables X and Y, one can always divide the computation in two steps by writing

$$\mathbb{E}(\varphi(X,Y)) = \mathbb{E}(\mathbb{E}(\varphi(X,Y)|Y)) = \mathbb{E}(\psi(Y))$$

where $\psi(y) = \mathbb{E}(\varphi(X,y))$ (this is actually nothing but Fubini's theorem).

Finally, the proposition below shows that Jensen's inequality also holds for conditional expectation.

Proposition 1.29. Let X be a random variable, $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$ and $\psi : \mathbb{R} \to \mathbb{R}$ be convex and such that $\mathbb{E}(|\psi(X)|) < \infty$. Then

$$\psi(\mathbb{E}(X|\mathcal{G})) \leq \mathbb{E}(\psi(X)|\mathcal{G}) \quad a.s.$$

In particular, $|\mathbb{E}(X|\mathcal{G})| \leq \mathbb{E}(|X||\mathcal{G})$ a.s.

Conditioning with respect to a generic random variable Y.

Once the definition of conditional expectation with respect to a σ-field is set, it is natural to define for a generic random variable Y:

$$\mathbb{E}(X|Y) = \mathbb{E}(X|\sigma(Y)) \quad \text{and} \quad \mathbb{P}(A|\mathcal{G}) = \mathbb{P}(A|\sigma(Y)).$$

Remark. Since any $\sigma(Y)$-measurable random variable may be written as $g(Y)$, where g is a Borel-measurable function, the definition of $\mathbb{E}(X|Y)$ may be rephrased as follows.

Definition 1.30. $E(X|Y) = \psi(Y)$, where $\psi : \mathbb{R} \to \mathbb{R}$ is the unique Borel-measurable function such that $\mathbb{E}(\psi(Y)\, g(Y)) = \mathbb{E}(X\, g(Y))$ for any function $g : \mathbb{R} \to \mathbb{R}$ Borel-measurable and bounded.

In two particular cases, the function ψ can be made explicit.

- As already seen above, if X, Y are two discrete random variables with values in a countable set C, then

$$E(X|Y) = \psi(Y), \quad \text{where} \quad \psi(y) = \sum_{x \in D} x\, \mathbb{P}(\{X = x\}|\{Y = y\}), \quad y \in C.$$

- If X, Y are two jointly continuous random variables with joint pdf $f_{X,Y}$, then

$$E(X|Y) = \psi(Y), \quad \text{where} \quad \psi(y) = \int_{\mathbb{R}} x\, \frac{f_{X,Y}(x,y)}{f_Y(y)}\, dy, \quad y \in \mathbb{R},$$

and f_Y is the marginal pdf of Y given by $f_Y(y) = \int_{\mathbb{R}} f_{X,Y}(x,y)\, dy$, assumed here to be strictly positive. Let us check that the random variable $\psi(Y)$ is indeed the conditional expectation of X given Y according to Definition 1.30: for any function $g : \mathbb{R} \to \mathbb{R}$ Borel-measurable and bounded, one has

$$\begin{aligned}
\mathbb{E}(\psi(Y)\, g(Y)) &= \int_{\mathbb{R}} \psi(y)\, g(y)\, f_Y(y)\, dy \\
&= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} x\, \frac{f_{X,Y}(x,y)}{f_Y(y)}\, dy \right) g(y)\, f_Y(y)\, dy \\
&= \iint_{\mathbb{R}^2} x\, g(y)\, f_{X,Y}(x,y)\, dx\, dy = \mathbb{E}(X\, g(Y)).
\end{aligned}$$