Chapter II

Probability Spaces

A probability space is a triplet $(\Omega, \mathcal{H}, \mathbb{P})$ where Ω is a set, $\mathcal{H}$ is a σ-algebra on Ω, and $\mathbb{P}$ is a probability measure on $(\Omega, \mathcal{H})$. Thus, mathematically, a probability space is a special measure space where the measure has total mass one.

But, our attitude and emotional response toward one is entirely different from those toward the other. On a measure space everything is deterministic and certain, on a probability space we face randomness and uncertainty.

A probability space $(\Omega, \mathcal{H}, \mathbb{P})$ is a mathematical model of a random experiment, an experiment whose exact outcome cannot be told in advance. The set Ω stands for the collection of all possible outcomes of the experiment. A subset H is said to *occur* if the outcome of the experiment happens to belong to H. Given our capabilities to measure, detect, and discern, and given the nature of answers we seek, only certain subsets H are distinguished enough to be of concern whether they occur. The σ-algebra $\mathcal{H}$ is the collection of all such subsets whose occurrence are noteworthy and decidable; the elements of $\mathcal{H}$ are called *events*. From this point of view, the conditions for $\mathcal{H}$ to be a σ-algebra are logical consequences of the interpretation of the term "event". Finally, for each event H, the chances that H occurs is modeled to be the number $\mathbb{P}(H)$, called the probability that H occurs.

The actual assignment of probabilities to events is the primary task of the probabilist. It requires much thought and experience, it is rarely explicit, and it determines the quality of the probability space as a model of the experiment involved. Once the probability space is fixed, the main task is to evaluate various integrals of interest by making adroit use of those implicitly defined probabilities. Often, the results are compared against experience, and the probability space is altered for a better fit.

Our aim in this chapter is to introduce the language and notation of probability theory. Implicit in the language are whole sets of attitudes, prejudices, and desires with which we hope to infect the reader.

E. Çınlar, *Probability and Stochastics*, Graduate Texts in Mathematics 261, DOI 10.1007/978-0-387-87859-1_2,

1 Probability Spaces and Random Variables

Let $(\Omega, \mathcal{H}, \mathbb{P})$ be a probability space. The set Ω is called the *sample space*; its elements are called *outcomes*. The σ-algebra $\mathcal{H}$ may be called the grand *history*; its elements are called *events*. We repeat the properties of the probability measure $\mathbb{P}$; all sets here are events:

1.1 *Norming:* $\mathbb{P}(\emptyset) = 0, \qquad \mathbb{P}(\Omega) = 1.$

Monotonicity: $H \subset K \quad \Rightarrow \quad \mathbb{P}(H) \leq \mathbb{P}(K).$

Finite additivity: $H \cap K = \emptyset \quad \Rightarrow \quad \mathbb{P}(H \cup K) = \mathbb{P}(H) + \mathbb{P}(K).$

Countable additivity: (H_n) disjointed $\quad \Rightarrow \quad \mathbb{P}(\bigcup_n H_n) = \sum_n \mathbb{P}(H_n).$

Sequential continuity: $H_n \nearrow H \quad \Rightarrow \quad \mathbb{P}(H_n) \nearrow \mathbb{P}(H),$
$H_n \searrow H \quad \Rightarrow \quad \mathbb{P}(H_n) \searrow \mathbb{P}(H).$

Boole's inequality: $\mathbb{P}(\bigcup_n H_n) \quad \leq \quad \sum_n \mathbb{P}(H_n).$

All of these are as before for arbitrary measures, except for the sequential continuity under decreasing limits, which is made possible by the finiteness of $\mathbb{P}$: If $H_1 \supset H_2 \supset \ldots$ and $\lim H_n = \bigcap H_n = H$, then the complements H_n^c increase to H^c, which implies that $\mathbb{P}(H_n^c) \nearrow \mathbb{P}(H^c)$ by the sequential continuity of measures under increasing limits, and we have $\mathbb{P}(H) = 1 - \mathbb{P}(H^c)$, and similarly for each H_n, by the finite additivity and norming of $\mathbb{P}$.

Negligibility, completeness

The concepts are the same as for arbitrary measures: A subset N of Ω is said to be *negligible* if there exists an event H such that $N \subset H$ and $\mathbb{P}(H) = 0$. The probability space is said to be *complete* if every negligible set is an event.

Improbable events do not bother the probabilist. Negligible sets should not either, but if a negligible set does not belong to $\mathcal{H}$ then we are not able to talk of its probability, which thing is bothersome. So, it is generally nicer to have $(\Omega, \mathcal{H}, \mathbb{P})$ complete. If it is not, it can be completed using Proposition I.3.10.

Almost surely, almost everywhere

An event is said to be *almost sure* if its probability is one. If a proposition holds for every outcome ω in an almost sure event, then we say that the proposition holds *almost surely* or *almost everywhere* or *for almost every* ω or *with probability one*. Obviously, the concept is equivalent to having the proposition fail only over a negligible set.

Random variables

Let $(E, \mathcal{E})$ be a measurable space. A mapping $X : \Omega \mapsto E$ is called a *random variable* taking values in $(E, \mathcal{E})$ provided that it be measurable relative to $\mathcal{H}$ and $\mathcal{E}$, that is, if

$$X^{-1}A = \{X \in A\} = \{\omega \in \Omega : X(\omega) \in A\} \tag{1.2}$$

is an event for every A in $\mathcal{E}$. Of course, it is sufficient to check the condition for A in a collection that generates $\mathcal{E}$. It is customary to denote random variables by capital letters.

If the σ-algebra $\mathcal{E}$ is understood from context, then we merely say that X takes values in E or that X is *E-valued*. This is especially the case if E is $\mathbb{R}$ or $\mathbb{R}^d$ or some Borel subset of some such space and $\mathcal{E}$ is the Borel σ-algebra on E.

The simplest random variables are indicators of events; we use the usual notation 1_H for the indicator of H. A random variable is *simple* if it takes only finitely many values, all in $\mathbb{R}$. It is said to be *discrete* if it is elementary, that is, if it takes only countably many values.

Distribution of a random variable

Let X be a random variable taking values in some measurable space $(E, \mathcal{E})$. Let μ be the image of $\mathbb{P}$ under X (see section I.5 for image measures), that is,

$$\mu(A) = \mathbb{P}(X^{-1}A) = \mathbb{P}\{X \in A\}, \qquad A \in \mathcal{E}, \tag{1.3}$$

where the last member is read as "the probability that X is in A". Then, μ is a probability measure on $(E, \mathcal{E})$; it is called the *distribution* of X.

In view of Proposition I.3.7, to specify the distribution μ, it is sufficient to specify $\mu(A)$ for all A belonging to a p-system that generates $\mathcal{E}$. In particular, if $E = \bar{\mathbb{R}}$ and $\mathcal{E} = \mathcal{B}_E$, the intervals $[-\infty, x]$ with x in $\mathbb{R}$ form a convenient p-system; consequently, in this case, it is enough to specify

$$c(x) = \mu[-\infty, x] = \mathbb{P}\{X \le x\}, \qquad x \in \mathbb{R}. \tag{1.4}$$

The resulting function $c : \mathbb{R} \mapsto [0, 1]$ is called the *distribution function* of X. Distribution functions are used extensively in elementary probability theory in order to avoid measures. We shall have little use for them. A review of some salient facts are put as exercises for the sake of completeness.

Functions of random variables

Let X be a random variable taking values in $(E, \mathcal{E})$. Let $(F, \mathcal{F})$ be another measurable space, and let $f : E \mapsto F$ be measurable relative to $\mathcal{E}$ and $\mathcal{F}$. Then, the composition $Y = f \circ X$ of X and f, namely,

$$Y(\omega) = f \circ X(\omega) = f(X(\omega)), \qquad \omega \in \Omega, \tag{1.5}$$

is a random variable taking values in $(F, \mathcal{F})$; this follows from Proposition I.2.5 that measurable functions of measurable functions are measurable. If μ is the distribution of X, then the distribution ν of Y is $\nu = \mu \circ f^{-1}$:

$$1.6 \qquad \nu(B) = \mathbb{P}\{Y \in B\} = \mathbb{P}\{X \in f^{-1}B\} = \mu(f^{-1}B), \qquad B \in \mathcal{F}.$$

Joint distributions

Let X and Y be random variables taking values in measurable spaces $(E, \mathcal{E})$ and $(F, \mathcal{F})$ respectively. Then, the pair $Z = (X, Y) : \omega \mapsto Z(\omega) = (X(\omega), Y(\omega))$ is measurable relative to $\mathcal{H}$ and the product σ-algebra $\mathcal{E} \otimes \mathcal{F}$, that is, Z is a random variable taking values in the product space $(E \times F, \mathcal{E} \otimes \mathcal{F})$.

The distribution of Z is a probability measure π on the product space and is also called the *joint distribution* of X and Y. Since $\mathcal{E} \otimes \mathcal{F}$ is generated by the p-system of measurable rectangles, in order to specify π it is sufficient to specify

$$1.7 \qquad \pi(A \times B) = \mathbb{P}\{X \in A, Y \in B\}, \qquad A \in \mathcal{E}, B \in \mathcal{F},$$

the right side being the probability that X is in A and Y is in B, that is, the probability of $\{X \in A\} \cap \{Y \in B\}$. In the opposite direction, given the joint distribution π, for A in $\mathcal{E}$ and B in $\mathcal{F}$, we have

$$1.8 \qquad \mu(A) = \mathbb{P}\{X \in A\} = \pi(A \times F), \qquad \nu(B) = \mathbb{P}\{Y \in B\} = \pi(E \times B).$$

In this context, the probability measures μ and ν are called the *marginal distributions* of X and Y respectively. These terms are used, with obvious generalizations, for any finite number of random variables.

Independence

Let X and Y be random variables taking values in $(E, \mathcal{E})$ and $(F, \mathcal{F})$ respectively, and let μ and ν be their respective (marginal) distributions. Then, X and Y are said to be *independent* if their joint distribution is the product measure formed by their marginals, that is, if the distribution of the pair (X, Y) is the product measure $\mu \times \nu$, or in still other words,

$$1.9 \qquad \mathbb{P}\{X \in A, Y \in B\} = \mathbb{P}\{X \in A\}\mathbb{P}\{Y \in B\}, \qquad A \in \mathcal{E}, B \in \mathcal{F}.$$

In probability theory, independence is used often as a primitive concept to be decided by considerations based on the underlying experiment and the way X and Y are defined. And, once it is decided upon, independence of X and Y becomes a convenient tool for specifying the joint distribution via its marginals. We shall return to these matters in Chapter IV for a rigorous treatment. For the present we mention an extension or two.

A finite collection $\{X_1, \ldots, X_n\}$ of random variables is said to be an *independency*, or the variables $X_1, \ldots, X_n$ are said to be *independent*, if the distribution of the random vector $(X_1, \ldots, X_n)$ has the product form $\mu_1 \times \cdots \times \mu_n$ where $\mu_1, \ldots, \mu_n$ are probability measures. Then, necessarily, μ_i is the distribution of X_i for each i. An arbitrary collection (countable or uncountable) of random variables is said to be an *independency* if every finite sub-collection of it is an independency.

Stochastic processes and probability laws

Let $(E, \mathcal{E})$ be a measurable space. Let T be an arbitrary set, countable or uncountable. For each t in T, let X_t be a random variable taking values in $(E, \mathcal{E})$. Then, the collection $\{X_t : t \in T\}$ is called a *stochastic process* with *state space* $(E, \mathcal{E})$ and *parameter set* T.

For each ω in Ω, let $X(\omega)$ denote the function $t \mapsto X_t(\omega)$ from T into E; then, $X(\omega)$ is an element of E^T. By Proposition I.6.27, the mapping $X : \omega \mapsto X(\omega)$ from Ω into E^T is measurable relative to $\mathcal{H}$ and $\mathcal{E}^T$. In other words, we may regard the stochastic process $\{X_t : t \in T\}$ as a random variable X that takes values in the product space $(F, \mathcal{F}) = (E^T, \mathcal{E}^T)$.

The distribution of the random variable X, that is, the probability measure $\mathbb{P} \circ X^{-1}$ on $(F, \mathcal{F})$, is called the *probability law* of the stochastic process $\{X_t : t \in T\}$.

Recall that the product σ-algebra $\mathcal{F}$ is generated by the finite-dimensional rectangles and, therefore, a probability measure on $(F, \mathcal{F})$ is determined by the values it assigns to those rectangles. It follows that the probability law of X is determined by the values

$$\mathbb{P}\{X_{t_1} \in A_1, \ldots, X_{t_n} \in A_n\} \tag{1.10}$$

with n ranging over $\mathbb{N}^*$, and $t_1, \ldots, t_n$ over T, and $A_1, \ldots, A_n$ over $\mathcal{E}$. Much of the theory of stochastic processes has to do with computing integrals concerning X from the given data regarding 1.10.

Examples of distributions

The aim here is to introduce a few distributions that are encountered often in probabilistic work. Other examples will appear in the exercises below and in the section next.

1.11 *Poisson distribution.* Let X be a random variable taking values in $\mathbb{N} = \{0, 1, \ldots\}$; it is to be understood that the relevant σ-algebra on $\mathbb{N}$ is the discrete σ-algebra of all subsets. Then, X is said to have the Poisson distribution with mean c if

$$\mathbb{P}\{X = n\} = \frac{e^{-c}\, c^n}{n!}\,, \qquad n \in \mathbb{N}.$$

Here, c is a strictly positive real number. The corresponding distribution is the probability measure μ on $\mathbb{N}$ defined by

$$\mu(A) = \sum_{n \in A} \frac{e^{-c} c^n}{n!}\,, \qquad A \subset \mathbb{N}.$$

1.12 *Exponential distributions.* Let X be a random variable with values in $\mathbb{R}_+$; the relevant σ-algebra on $\mathbb{R}_+$ is $\mathcal{B}(\mathbb{R}_+)$. Then, X is said to have the exponential distribution with scale parameter c if its distribution μ has the form

$$\mu(dx) = dx\; ce^{-cx}\,, \quad x \in \mathbb{R}_+,$$

where dx is short for $\mathrm{Leb}(dx)$. Here, $c > 0$ is a constant, and we used the form I.5.8 to display μ. In other words, μ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}_+$ and its density function is $p(x) = ce^{-cx}$, $x \in \mathbb{R}_+$. When $c = 1$, this distribution is called the standard exponential.

1.13 *Gamma distributions.* Let X be a random variable with values in $\mathbb{R}_+$. It is said to have the gamma distribution with shape index a and scale parameter c if its distribution μ has the form

$$\mu(dx) = dx\; \frac{c^a x^{a-1} e^{-cx}}{\Gamma(a)}\,, \qquad x \in \mathbb{R}_+.$$

Here, $a > 0$ and $c > 0$ are constants and $\Gamma(a)$ is the so-called gamma function. The last is defined so that μ is a probability measure, that is,

$$\Gamma(a) = \int_0^\infty dx\; x^{a-1} e^{-x}.$$

Incidentally, the density function for μ takes the value $+\infty$ at $x = 0$ if $a < 1$, but this is immaterial since $\mathrm{Leb}\{0\} = 0$; or, in probabilistic terms, $X \in \mathbb{R}_+^* = (0, \infty)$ almost surely, and it is sufficient to define the density on $\mathbb{R}_+^*$. In general, $\Gamma(a) = (a-1)\Gamma(a-1)$ for $a > 1$. This allows one, together with $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ and $\Gamma(1) = 1$, to give an explicit expression for $\Gamma(a)$ when $a > 0$ is an integer or half-integer. In particular, when $a = 1$, the gamma distribution becomes the exponential; and when $c = \frac{1}{2}$ and $a = \frac{n}{2}$ for some integer $n \geq 1$, it is also called the Chi-square distribution with n degrees of freedom. Finally, when $c = 1$, we call the distribution standard gamma distribution with shape index a.

1.14 *Gaussian distributions.* Let X be a real-valued random variable. It is said to have the Gaussian (or normal) distribution with mean a and variance b if its distribution μ has the form

$$\mu(dx) = dx\; \frac{1}{\sqrt{2\pi b}}\, e^{-(x-a)^2/2b}, \qquad x \in \mathbb{R}.$$

Here, $a \in \mathbb{R}$ and $b > 0$, both constant. If $a = 0$ and $b = 1$, then μ is called the standard Gaussian distribution.

1.15 *Independent gamma variables.* Let γ_a denote the standard gamma distribution with shape index a; this is the probability measure μ of Example 1.13 above but with $c = 1$. Let X have the distribution γ_a, and Y the distribution γ_b; here $a > 0$ and $b > 0$. Suppose that X and Y are independent. Then, the joint distribution of X and Y is the product measure $\gamma_a \times \gamma_b$, that is, the distribution of the pair (X, Y) is the probability measure π on $\mathbb{R}_+ \times \mathbb{R}_+$ given by

$$\pi(dx, dy) = \gamma_a(dx)\ \gamma_b(dy) = dx\ dy\ \frac{e^{-x}\ x^{a-1}}{\Gamma(a)} \cdot \frac{e^{-y}\ y^{b-1}}{\Gamma(b)}.$$

1.16 *Gaussian with exponential variance.* Let X and Y be random variables taking values in $\mathbb{R}_+$ and $\mathbb{R}$ respectively. Suppose that their joint distribution π is given by

$$\pi(dx, dy) = dx\ dy\ ce^{-cx}\ \frac{1}{\sqrt{2\pi x}}\ e^{-y^2/2x}\ , \qquad x \in \mathbb{R}_+,\ y \in \mathbb{R}.$$

Note that π has the form $\pi(dx, dy) = \mu(dx)\ K(x, dy)$, where μ is the exponential distribution with scale parameter c, and for each x, the distribution $B \mapsto K(x, B)$ is Gaussian with mean 0 and variance x. Indeed, K is a transition kernel from $\mathbb{R}_+$ into $\mathbb{R}$, and π is an instance of the measure appearing in Theorem I.6.11. It is clear that the marginal distribution of X is the exponential distribution μ. The marginal distribution ν of Y has the form $\nu = \mu K$ introduced in Theorem I.6.3:

$$\nu(B) = \pi(\mathbb{R}_+ \times B) = \int_{\mathbb{R}_+} \mu(dx)\ K(x, B)\ , \qquad B \in \mathcal{B}_{\mathbb{R}}.$$

It is seen easily that ν is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$, that is, ν has the form $\nu(dy) = dy \cdot n(y)$, and the density function is

$$n(y) = \int_0^\infty dx\ ce^{-cx}\ \frac{e^{-y^2/2x}}{\sqrt{2\pi x}} = \tfrac{1}{2}\ b\ e^{-b|y|}\ , \qquad y \in \mathbb{R},$$

with $b = \sqrt{2c}$. Incidentally, this distribution ν is called the *two-sided exponential* distribution with parameter b. Finally, we note that π is not the product $\mu \times \nu$, that is, X and Y are dependent variables.

Exercises and complements

1.17 *Distribution functions.* Let X be a random variable taking values in $\bar{\mathbb{R}} = [-\infty, +\infty]$. Let μ be its distribution, and c its distribution function, defined by 1.4. Then, c is a function from $\mathbb{R}$ into $[0, 1]$. It is increasing and right-continuous as indicated in Exercise I.5.14.

a) Since c is increasing, the left-hand limit

$$c(x-) = \lim_{y \uparrow x} c(y)$$

exists for every x in $\mathbb{R}$. Similarly, the limits

$$c(-\infty) = \lim_{x \downarrow -\infty} c(x) \qquad c(+\infty) = \lim_{x \uparrow \infty} c(x)$$

exist. Show that

$$c(x-) = \mathbb{P}\{X < x\}, \qquad c(x) - c(x-) = \mathbb{P}\{X = x\}$$

$$c(-\infty) = \mathbb{P}\{X = -\infty\}, \qquad c(+\infty) = \mathbb{P}\{X < \infty\} = 1 - \mathbb{P}\{X = \infty\}.$$

b) Let D be the set of all atoms of the distribution μ. Then, D consists of all x in $\mathbb{R}$ for which $c(x) - c(x-) > 0$, plus the point $-\infty$ if $c(-\infty) > 0$, plus the point $+\infty$ if $c(\infty) < 1$. Of course, D is countable. Define $D_x = D \cap (-\infty, x]$ and

$$a(x) = c(-\infty) + \sum_{y \in D_x} [c(y) - c(y-)] \,, \qquad b(x) = c(x) - a(x)$$

for x in $\mathbb{R}$. Then, a is an increasing right-continuous function that increases by jumps only, and b is increasing continuous. Show that a is the distribution function of the measure

$$\mu_a(B) = \mu(B \cap D) \,, \qquad B \in \mathcal{B}(\bar{\mathbb{R}}),$$

and b is the distribution function of the measure $\mu_b = \mu - \mu_a$. Note that μ_a is purely atomic and μ_b is diffuse. The random variable X is almost surely discrete if and only if $\mu = \mu_a$, that is, $a = c$.

1.18 *Quantile functions.* Let X be real-valued, let c be its distribution function. Note that, then, $c(-\infty) = 0$ and $c(+\infty) = 1$. Suppose that c is continuous and strictly increasing, and let q be the functional inverse of c, that is, $q(u) = x$ if and only if $c(x) = u$ for u in $(0, 1)$. The function $q : \ (0, 1) \mapsto \mathbb{R}$ is called the quantile function of X since

$$\mathbb{P}\{X \leq q(u)\} = u \,, \qquad u \in (0, 1).$$

Let U be a random variable having the *uniform distribution* on $(0, 1)$, that is, the distribution of U is the Lebesgue measure on $(0, 1)$. Show that, then, the random variable $Y = q \circ U$ has the same distribution as X. In general, $Y \neq X$.

1.19 *Continuation.* This is to re-do the preceding exercise assuming that $c : \mathbb{R} \mapsto [0, 1]$ is only increasing and right-continuous. Let $q : \ (0, 1) \mapsto \bar{\mathbb{R}}$ be the right-continuous functional inverse of c, that is,

$$q(u) = \inf\{x \in \mathbb{R} : \quad c(x) > u\}$$

with the usual conventions that $\inf \mathbb{R} = -\infty$, $\inf \emptyset = +\infty$. We call q the quantile function corresponding to c by analogy with the preceding exercise. Recall from Exercise I.5.13 that q is increasing and right-continuous, and that c is related to q by the same formula with which q is related to c. Note that q is real-valued if and only if $c(-\infty) = 0$ and $c(+\infty) = 1$. See also Figure 1. Show that $c(x-) \le u$ if and only if $q(u) \ge x$, and, by symmetry, $q(u-) \le x$ if and only if $c(x) \ge u$.

1.20 *Construction of probability measures on* $\bar{\mathbb{R}}$. Let c be a cumulative distribution function, that is, $c : \mathbb{R} \mapsto [0,1]$ is increasing and right-continuous. Let $q : (0,1) \mapsto \bar{\mathbb{R}}$ be the corresponding quantile function. Let λ denote the Lebesgue measure on $(0,1)$ and put $\mu = \lambda \circ q^{-1}$. Show that μ is a probability measure on $\bar{\mathbb{R}}$. Show that μ is the distribution on $\bar{\mathbb{R}}$ corresponding to the distribution function c. Thus, to every distribution function c on $\mathbb{R}$ there corresponds a unique probability measure μ on $\bar{\mathbb{R}}$ and vice-versa.

1.21 *Construction of random variables.* Let μ be a probability measure on $\bar{\mathbb{R}}$. Then, there exists a probability space $(\Omega, \mathcal{H}, \mathbb{P})$ and a random variable $X : \Omega \mapsto \bar{\mathbb{R}}$ such that μ is the distribution of X: Take $\Omega = (0,1)$, $\mathcal{H} = \mathcal{B}_{(0,1)}$, $\mathbb{P} = \text{Leb}$, and define $X(\omega) = q(\omega)$ for ω in Ω, where q is the quantile function corresponding to the measure μ (via the cumulative distribution function). See Exercise I.5.15 for the extension of this construction to abstract spaces. This setup is the theoretical basis of Monte-Carlo studies.

1.22 *Supplement on quantiles.* Literature contains definitions similar to that in 1.19 for q, but with slight differences, one of the popular ones being

$$p(u) = \inf\{x \in \mathbb{R} : \quad c(x) \ge u\}\,, \qquad u \in (0,1).$$

Some people prefer supremums, but there is nothing different, since $q(u) = \sup\{x : \ c(x) \le u\}$ and $p(u) = \sup\{x : \ c(x) < u\}$. In fact, there is close relationship between p and q: we have $p(u) = q(u-) = \lim_{v \nearrow u} q(v)$. The function q is right-continuous, whereas p is left-continuous. We prefer q over p, because q and c are functional inverses of each other. Incidentally, in the constructions of 1.20 and 1.21 above, the minor difference between p and q proves unimportant: Since q is increasing and right-continuous, $p(u) = q(u-)$ differs from $q(u)$ for at most countably many u; therefore, $\text{Leb}\{u : p(u) \ne q(u)\} = 0$ and, hence, $\lambda \circ q^{-1} = \lambda \circ p^{-1}$ with $\lambda = \text{Leb}$ on $(0,1)$.

2 Expectations

Throughout this section $(\Omega, \mathcal{H}, \mathbb{P})$ is a probability space and all random variables are defined on Ω and take values in $\bar{\mathbb{R}}$, unless stated otherwise.

Let X be a random variable. Since it is $\mathcal{H}$-measurable, its integral with respect to the measure $\mathbb{P}$ makes sense to talk about. That integral is called the *expected value* of X and is denoted by any of the following

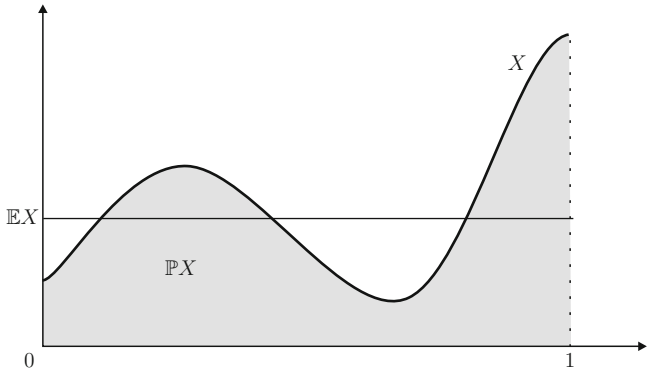

Figure 2: The integral $\mathbb{P}X$ is the area under X, the expected value $\mathbb{E}X$ is the constant "closest" to X.

2.1 $$\mathbb{E}X = \int_\Omega \mathbb{P}(d\omega)\, X(\omega) = \int_\Omega X\, d\mathbb{P} = \mathbb{P}X.$$

The expected value $\mathbb{E}X$ exists if and only if the integral does, that is, if and only if we do not have $\mathbb{E}X^+ = \mathbb{E}X^- = +\infty$. Of course, $\mathbb{E}X$ exists whenever $X \geq 0$, and $\mathbb{E}X$ exists and is finite if X is bounded.

We shall treat $\mathbb{E}$ as an operator, the expectation operator corresponding to $\mathbb{P}$, and call $\mathbb{E}X$ the expectation of X from time to time. The change in notation serves to highlight the important change in our interpretation of $\mathbb{E}X$: The integral $\mathbb{P}X$ is the "area under the function" X in a generalized sense. The expectation $\mathbb{E}X$ is the "weighted average of the values" of X, the weight distribution being specified by $\mathbb{P}$, the total weight being $\mathbb{P}(\Omega) = 1$. See Figure 2 above for the distinction.

Except for this slight change in notation, all the conventions and notations of integration are carried over to expectations. In particular, X is said to be *integrable* if $\mathbb{E}X$ exists and is finite. The integral of X over an event H is $\mathbb{E}X1_H$. As before with integrals, we shall state most results for positive random variables, because expectations exist always for such, and because the extensions to arbitrary random variables are generally obvious.

Properties of expectation

The following is a rapid summary of the main results on integrals stated in probabilistic terms. Here, X, Y, etc. are random variables taking values in $\bar{\mathbb{R}}$, and a, b, etc. are positive constants.

2.2 *Positivity:* $X \geq 0 \;\Rightarrow\; \mathbb{E}X \geq 0$.

Monotonicity: $X \geq Y \geq 0 \;\Rightarrow\; \mathbb{E}X \geq \mathbb{E}Y$.

Linearity: $X, Y \geq 0 \;\Rightarrow\; \mathbb{E}(aX + bY) = a\mathbb{E}X + b\mathbb{E}Y$.

Insensitivity: $X, Y \geq 0,\ X = Y$ almost surely $\Rightarrow\ \mathbb{E}X = \mathbb{E}Y$.

Monotone convergence: $X_n \geq 0,\ X_n \nearrow X\ \Rightarrow\ \mathbb{E}X_n \nearrow \mathbb{E}X$
$X_n \geq 0,\ \Rightarrow\ \mathbb{E}\sum X_n = \sum \mathbb{E}X_n$.

Fatou's Lemma: $X_n \geq 0\ \Rightarrow\ \mathbb{E}\liminf X_n\ \leq\ \liminf \mathbb{E}X_n$.

Dominated convergence: $|X_n| \leq Y,\ Y$ integrable, $\lim X_n$ exists
$\Rightarrow\ \mathbb{E}\lim X_n = \lim \mathbb{E}X_n$.

Bounded convergence: $|X_n| \leq b,\ b < \infty,\ \lim X_n$ exists
$\Rightarrow\ \mathbb{E}\lim X_n = \lim \mathbb{E}X_n$.

2.3 Remarks. a) Positivity can be added to: for $X \geq 0$, we have $\mathbb{E}X = 0$ if and only if $X = 0$ almost surely.

b) Monotonicity can be extended: if $X \geq Y$, then $\mathbb{E}\,X \geq \mathbb{E}\,Y$ provided that both $\mathbb{E}\,X$ and $\mathbb{E}\,Y$ exist (infinite values are allowed); if $X \geq Y$ and either $\mathbb{E}X$ or $\mathbb{E}Y$ is finite, then both $\mathbb{E}X$ and $\mathbb{E}Y$ exist and $\mathbb{E}X \geq \mathbb{E}Y$.

c) Insensitivity can be extended likewise: if $X = Y$ almost surely and either $\mathbb{E}X$ or $\mathbb{E}Y$ exists, then so is the other and $\mathbb{E}X = \mathbb{E}Y$.

d) The preceding two remarks have a useful partial converse: If $\mathbb{E}\ X1_H \geq \mathbb{E}\ Y1_H$ for every event H, then $X \geq Y$ almost surely. To show this, we use the remark above on monotonicity and the assumed inequality with $H = \{X < q < r < Y\}$, where q and r are rational numbers with $q < r$. This yields

$$q\mathbb{P}(H) = \mathbb{E}\ q1_H \geq \mathbb{E}\ X1_H \geq \mathbb{E}\ Y1_H \geq \mathbb{E}\ r1_H = r\mathbb{P}(H),$$

which is possible with $q < r$ only if $\mathbb{P}(H) = 0$. Hence, the event $\{Y > X\}$ has probability zero, via Boole's inequality, since it is the union of events like H over all rationals q and r with $q < r$.

e) Convergence theorems have various generalizations along the lines indicated for integrals. For example, an easy consequence of the monotone convergence theorem is that if $X_n \leq Y$ for all n for some integrable Y, and if $X_n \searrow X$, then $\mathbb{E}X_n \searrow \mathbb{E}X$.

f) Convergence theorems have almost sure versions similar to almost everywhere versions with integrals.

g) If a mapping $X : \Omega \mapsto \bar{\mathbb{R}}$ is equal to a random variable Y almost surely, and even if $X(\omega)$ is specified only for almost every ω, the expected value of X is defined to be $\mathbb{E}Y$.

Expectations and integrals

The following relates expectations, which are integrals with respect to $\mathbb{P}$, to integrals with respect to distributions. This is the work horse of computations. Recall that $\mathcal{E}_+$ is the collection of all positive $\mathcal{E}$-measurable functions (from E into $\bar{\mathbb{R}}_+$).

2.4 THEOREM. *Let X be a random variable taking values in some measurable space $(E, \mathcal{E})$. If μ is the distribution of X, then*

$$\mathbb{E}\ f\circ X = \mu f \tag{2.5}$$

for every f in $\mathcal{E}_+$. Conversely, if 2.5 holds for some measure μ and all f in $\mathcal{E}_+$, then μ is the distribution of X.

Proof. The first statement is a re-phrasing of Theorem I.5.2 on integration with respect to image measures: if $\mu = \mathbb{P}\circ X^{-1}$, then $\mu f = \mathbb{P}(f\circ X) = \mathbb{E}\ f\circ X$ at least for f in $\mathcal{E}_+$. Conversely, if 2.5 holds for all f in $\mathcal{E}_+$, taking $f = 1_A$ in particular, we see that

$$\mu(A) \ = \ \mu\ 1_A \ = \ \mathbb{E}\ 1_A \circ X \ = \ \mathbb{P}\{X \in A\},$$

that is, μ is the distribution of X. □

In the preceding theorem, the restriction to positive f is for reasons of convenience. For f in $\mathcal{E}$, the formula 2.5 holds for f^+ and f^- respectively, and hence for f, provided that either the expectation $\mathbb{E}\ f\circ X$ or the integral μf exists (then so does the other). The converse statement is useful for figuring out the distribution of X in cases where X is a known function of other random variables whose joint distribution is known. In such cases, it encompasses the formula 1.6 and is more intuitive; we shall see several illustrations of its use below.

Obviously, for a measure μ to be the distribution of X it is sufficient to have 2.5 hold for all f having the form $f = 1_A$ with A in $\mathcal{E}$, or with A in some p-system generating $\mathcal{E}$. When E is a metrizable topological space and $\mathcal{E} = \mathcal{B}(E)$, it is also sufficient to have 2.5 hold for all f that are bounded, positive, and continuous; see Exercise 2.36 in this connection.

Means, variances, Laplace and Fourier transforms

Certain expected values have special names. Let X be a random variable taking values in $\bar{\mathbb{R}}$ and having the distribution μ. The expected value of the n^{th} power of X, namely $\mathbb{E}X^n$, is called the n^{th} *moment* of X. In particular, $\mathbb{E}X$ is also called the *mean* of X. Assuming that the mean is finite (that is, X is integrable), say $\mathbb{E}X = a$, the n^{th} moment of $X - a$ is called the n^{th} centered moment of X. In particular, $\mathbb{E}(X-a)^2$ is called the *variance* of X, and we shall denote it by $\text{Var}X$; note that

$$\text{Var}\ X = \mathbb{E}\ (X-a)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2, \tag{2.6}$$

assuming of course that $a = \mathbb{E}X$ is finite.

Assuming that X is positive, for r in $\mathbb{R}_+$, the random variable e^{-rX} takes values in the interval $[0,1]$, and its expectation

$$\hat{\mu}_r = \mathbb{E}\ e^{-rX} = \int_{\mathbb{R}_+} \mu(dx)\ e^{-rx} \tag{2.7}$$

is a number in $[0,1]$. The resulting function $r \mapsto \hat{\mu}_r$ from $\mathbb{R}_+$ into $[0,1]$ is called the *Laplace transform* of the distribution μ, and by an abuse of language, also the Laplace transform of X.

It can be shown that the Laplace transform determines the distribution: if μ and ν are distributions on $\bar{\mathbb{R}}_+$, and $\hat{\mu}_r = \hat{\nu}_r$ for all r in $\mathbb{R}_+$, then $\mu = \nu$: see Exercise 2.36 below.

Suppose that X is real-valued, that is, X takes values in $\mathbb{R}$. For r in $\mathbb{R}$, $e^{irX} = \cos rX + i \sin rX$ is a complex-valued random variable (here $i = \sqrt{-1}$), and the notion of expected value extends to it naturally:

$$2.8 \qquad \hat{\mu}_r = \mathbb{E}\, e^{irX} = \int_{\mathbb{R}} \mu(dx)\, e^{irx} = \int_{\mathbb{R}} \mu(dx)\, \cos rx + i \int_{\mathbb{R}} \mu(dx)\, \sin rx.$$

The resulting complex-valued function $r \mapsto \hat{\mu}_r$ from $\mathbb{R}$ into the complex plane is called the *Fourier transform* of the distribution μ, or the *characteristic function* of the random variable X. As with Laplace transforms, the Fourier transform determines the distribution.

Finally, if X takes values in $\bar{\mathbb{N}} = \{0, 1, \ldots, +\infty\}$, then

$$2.9 \qquad \mathbb{E}\, z^X = \sum_{n=0}^{\infty} z^n\, \mathbb{P}\{X = n\}\,, \qquad z \in [0,1],$$

defines a function from $[0,1]$ into $[0,1]$ which is called the *generating function* of X. It determines the distribution of X: in a power series expansion of it, the coefficient of z^n is $\mathbb{P}\{X = n\}$ for each n in $\mathbb{N}$.

Examples

2.10 *Gamma distribution.* Fix $a > 0$ and $c > 0$, and let $\gamma_{a,c}$ be the gamma distribution with shape index a and scale parameter c; see Example 1.13. Let X have $\gamma_{a,c}$ as its distribution. Then, X has finite moments of all orders. Indeed, for every p in $\mathbb{R}_+$,

$$\begin{aligned} \mathbb{E}\, X^p &= \int_0^\infty \gamma_{a,c}(dx)\, x^p = \int_0^\infty dx\, \frac{c^a\, x^{a-1}\, e^{-cx}}{\Gamma(a)}\, x^p \\ &= \frac{\Gamma(a+p)}{c^p\, \Gamma(a)} \int_0^\infty dx\, \frac{c^{a+p}\, x^{a+p-1}\, e^{-cx}}{\Gamma(a+p)} = \frac{\Gamma(a+p)}{\Gamma(a)}\, c^{-p}, \end{aligned}$$

since the last integral is $\gamma_{a+p,c}(\mathbb{R}_+) = 1$. Finally, to explain the term "scale parameter" for c, we show that cX has the standard gamma distribution with shape index a (to understand the term "shape index" draw the density function of γ_a for $a < 1, a = 1, a > 1$). To this end, we use Theorem 2.4. Let f be a positive Borel function on $\mathbb{R}_+$. Then,

$$\mathbb{E} f(cX) = \int_0^\infty dx\, \frac{c^a\, x^{a-1}\, e^{-cx}}{\Gamma(a)}\, f(cx) = \int_0^\infty dy\, \frac{y^{a-1}\, e^{-y}}{\Gamma(a)}\, f(y),$$

which means that cX has the distribution γ_a, the standard gamma with shape index a.

2.11 *Gamma and gamma and beta.* Let X and Y be as in Example 1.15, that is, X and Y are independent, X has the standard gamma distribution γ_a with shape index a, and Y has the standard gamma distribution γ_b with shape index b. We now show that

a) $X + Y$ has the standard gamma distribution γ_{a+b},
b) $X/(X + Y)$ has the distribution

$$\beta_{a,b}(du) = du\, \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, u^{a-1}\, (1-u)^{b-1}\,, \quad 0 < u < 1,$$

which is called the beta distribution with index pair (a, b), and

c) $X + Y$ and $X/(X + Y)$ are independent, that is, their joint distribution π is the product measure $\gamma_{a+b} \times \beta_{a,b}$.

We show all this by using the method of Theorem 2.4. Let f be a positive Borel function on $\mathbb{R}_+ \times [0, 1]$ and consider the integral πf:

$$\begin{aligned}
\pi f &= \mathbb{E}\, f(X+Y, \frac{X}{X+Y}) \\
&= \int_0^\infty dx\, \frac{x^{a-1}\, e^{-x}}{\Gamma(a)} \int_0^\infty dy\, \frac{y^{b-1}\, e^{-y}}{\Gamma(b)}\, f(x+y, \frac{x}{x+y}) \\
&= \int_0^\infty dz \int_0^1 du\, \frac{z^{a+b-1}\, e^{-z}}{\Gamma(a)\, \Gamma(b)}\, u^{a-1}\, (1-u)^{b-1}\, f(z,u),
\end{aligned}$$

where the last line is obtained by replacing x with uz and y with $(1-u)z$, and noting that the Jacobian of the transformation is equal to z. There remains to note that the last expression is equal to $(\gamma_{a+b} \times \beta_{a,b})f$, which proves all three claims together.

2.12 *Laplace transforms and distributions and Pareto.* Let X be a random variable taking values in $\mathbb{R}_+$. Then, the Laplace transform $r \mapsto \mathbb{E}\, e^{-rX}$ is a decreasing continuous function on $\mathbb{R}_+$ with value 1 at $r = 0$. Hence, there is a positive random variable R such that

$$\mathbb{P}\{R > r\} = \mathbb{E}\, e^{-rX}, \quad r \in \mathbb{R}_+.$$

We now show that, in fact, we may take

$$R = Y/X,$$

where Y is independent of X and has the standard exponential distribution: Letting μ denote the distribution of X, for r in $\mathbb{R}_+$,

$$\begin{aligned}
\mathbb{P}\{R > r\} &= \mathbb{P}\{Y > rX\} \\
&= \int_{\mathbb{R}_+} \mu(dx) \int_{\mathbb{R}_+} dy\, e^{-y}\, 1_{(rx,\infty)}(y) \\
&= \int_{\mathbb{R}_+} \mu(dx)\, e^{-rx} = \mathbb{E}\, e^{-rX}
\end{aligned}$$

as was to be shown. In particular, if X has the gamma distribution with shape index a and scale c, then

$$\mathbb{P}\{R > r\} = \mathbb{E}\, e^{-rX} = \left(\frac{c}{c+r}\right)^a, \quad r \in \mathbb{R}_+,$$

according to the Laplace transform computation above in 2.11. Then, R is said to have the *Pareto* distribution with shape index a and scale parameter c. Since $R = Y/X$ and X is "small" in the sense that all its moments are finite, R should be big in the sense that its distribution should have a heavy tail.

Exercises and complements

Some of these are re-statements of results on integrals served up in probabilistic terms. Some are elementary facts that are worth recalling. And some are useful complements. Throughout, X, Y, etc. are random variables.

2.13 *Finiteness.* If $X \geq 0$ and $\mathbb{E}X < \infty$, then $X < \infty$ almost surely. More generally, if X is integrable then it is real-valued almost surely. Show.

2.14 *Moments of positive variables.* If $X \geq 0$, then for every p in $\mathbb{R}_+$,

$$\mathbb{E}X^p = \int_0^\infty dx\ p\ x^{p-1}\ \mathbb{P}\{X > x\}.$$

Show this, using Fubini's theorem with the product measure $\mathbb{P} \times \text{Leb}$, after noting that

$$X^p(\omega) = \int_0^{X(\omega)} dx\ px^{p-1} = \int_0^\infty dx\ px^{p-1}\ 1_{\{X>x\}}(\omega).$$

In particular, if X takes values in $\bar{\mathbb{N}} = \{0, 1, \ldots, +\infty\}$, then

$$\mathbb{E}X = \sum_{n=0}^\infty \mathbb{P}\{X > n\}, \qquad \mathbb{E}X^2 = 2\sum_{n=0}^\infty n\ \mathbb{P}\{X > n\} + \mathbb{E}\ X.$$

2.15 *Optimality of* $\mathbb{E}X$. Define

$$f(a) = \int_\Omega \mathbb{P}(d\omega)\ (X(\omega) - a)^2, \quad a \in \mathbb{R},$$

that is, $f(a)$ is the "weighted sum of errors squared" if X is estimated to be the constant a. Show that f is minimized by $a = \mathbb{E}X$ and that the minimum value is Var X.

2.16 *Variance.* Suppose that X is integrable. Show that, for a and b in $\mathbb{R}$,

$$\text{Var}\ (a + bX)\ =\ b^2\text{Var}\ X.$$

2.17 *Markov's inequality.* For $X \geq 0$,

$$\mathbb{P}\{X > b\} \leq \frac{1}{b}\, \mathbb{E}X$$

for every $b > 0$. Show this by noting that $X \geq b\ 1_{\{X>b\}}$.

2.18 *Chebyshev's inequality.* Suppose that X has finite mean. Apply Markov's inequality to $(X - \mathbb{E}X)^2$ to show that

$$\mathbb{P}\{|X - \mathbb{E}X| > \varepsilon\} \leq \frac{1}{\varepsilon^2}\ \mathrm{Var}\ X, \quad \varepsilon > 0.$$

2.19 *Markov's inequality generalized.* Let X be real-valued. Let $f : \mathbb{R} \mapsto \mathbb{R}_+$ be increasing. Show that, for every b in $\mathbb{R}$,

$$\mathbb{P}\{X > b\} \leq \frac{1}{f(b)}\ \mathbb{E}\ f \circ X.$$

2.20 *Jensen's inequality.* Let X have finite mean. Let f be a convex function on $\mathbb{R}$, that is, $f = \sup f_n$ for some sequence of functions f_n having the form $f_n(x) = a_n + b_n x$. Show that

$$\mathbb{E}\ f(X)\ \geq\ f(\mathbb{E}X).$$

2.21 *Gamma distribution.* This is to generalize Example 2.12 slightly by the use of the remark on "scale parameter" in Example 2.10. Let X and Y be independent, let X have distribution $\gamma_{a,c}$ and Y the distribution $\gamma_{b,c}$. Then, show that, $X + Y$ has the distribution $\gamma_{a+b,c}$, and $X/(X + Y)$ has the same old distribution $\beta_{a,b}$, and the two random variables are independent.

2.22 *Gaussian variables.* Show that X has the Gaussian distribution with mean a and variance b if and only if $X = a + \sqrt{b}Z$ for some random variable Z that has the standard Gaussian distribution. Show that

$$\begin{aligned} &\mathbb{E}\ Z = 0, \quad \mathrm{Var}\ Z = 1, \quad \mathbb{E}\ e^{irZ} = e^{-r^2/2}, \\ &\mathbb{E}\ X = a, \quad \mathrm{Var}\ X = b, \quad \mathbb{E}\ e^{irX} = e^{ira - r^2 b/2}. \end{aligned}$$

2.23 *Gamma-Gaussian connection.* a) Let Z have the Gaussian distribution with mean 0 and variance b. Show that, then, $X = Z^2$ has the gamma distribution with shape index $a = 1/2$ and scale parameter $c = 1/2b$. Hint: Compute $\mathbb{E}\ f \circ X = \mathbb{E}\ g \circ Z$ with $g(z) = f(z^2)$ and use Theorem 2.4 to identify the result.

b) Let $Z_1, \ldots, Z_n$ be independent standard Gaussian variables. Show that the sum of their squares has the gamma distribution with shape index $n/2$ and scale $1/2$.

2.24 *Uniform distribution.* Let $a < b$ be real numbers. Uniform distribution on (a, b) is the Lebesgue measure on (a, b) normalized to have mass one, that is, $\frac{1}{b-a}$Leb. The standard case is where $a = 0$ and $b = 1$. Since the Lebesgue measure puts no mass at points, the uniform distribution on $[a, b]$ is practically the same as that on (a, b). Let U have the standard uniform distribution on $(0, 1)$; let q be a quantile function. Then $q \circ U$ is a random variable having q as its quantile function.

2.25 *Uniform and exponential.* Let U have the uniform distribution on $(0, 1)$. Let $X = -\frac{1}{c} \log U$. Show that X has the exponential distribution with scale parameter c.

2.26 *Exponential-Gaussian-Uniform.* Let U and V be independent and uniformly distributed on $(0, 1)$. Let $R = \sqrt{-2 \log U}$, so that R^2 has the exponential distribution with scale parameter $1/2$, that is, R^2 has the same distribution as the sum of the squares of two independent standard Gaussian variables.Define

$$X = R \cos 2\pi V, \qquad Y = R \sin 2\pi V.$$

Show that X and Y are independent standard Gaussian variables. Show that, conversely, if X and Y are independent standard Gaussian variables, then the polar coordinates R and A of the random point (X, Y) in $\mathbb{R}^2$ are independent, R^2 has the exponential distribution with scale parameter $1/2$, and A has the uniform distribution on $[0, 2\pi]$.

2.27 *Cauchy distribution.* Let X and Y be independent standard Gaussian variables. Show that the distribution μ of $Z = X/Y$ has the form

$$\mu(dz) = dz \, \frac{1}{\pi(1 + z^2)}, \qquad z \in \mathbb{R}.$$

It is called the Cauchy distribution. Note that, if a random variable Z has the Cauchy distribution, then so does $1/Z$. Also, show that, if A has the uniform distribution on $(0, 2\pi)$, then $\tan A$ and $\cot A$ are both Cauchy distributed.

2.28 *Sums and transforms.* Let X and Y be independent positive random variables. Show that the Laplace transform for $X + Y$ is the product of the Laplace transforms for X and Y. Since the Laplace transform of a distribution determines the distribution, this specifies the distribution of $X + Y$, at least in principle. When X and Y are real-valued (instead of being positive), the same statements hold for characteristic functions.

2.29 *Characteristic functions.* Let X and Y be independent gamma distributed random variables with respective shape indices a and b, and the same scale parameter c. Compute the characteristic functions of X, Y, $X+Y$, $X-Y$. Note, in particular, that $X+Y$ has the gamma distribution with shape index $a + b$ and scale c.

2.30 *Gaussian with gamma variance.* Let X and Y be independent, X having the gamma distribution $\gamma_{a,c}$ (with shape index a and scale parameter c), and Y having the standard Gaussian distribution. Recall that $\sqrt{b}Y$ has the Gaussian distribution with mean 0 and variance $b > 0$. We now replace b with X: let $Z = \sqrt{X}\, Y$. Show that

$$\mathbb{E}\, e^{irZ} = \mathbb{E}\, e^{-r^2X/2} = \left(\frac{2c}{2c+r^2}\right)^a, \quad r \in \mathbb{R}.$$

Let U and V be independent with the distribution $\gamma_{a,\sqrt{2c}}$ for both. Show that

$$\mathbb{E}\, e^{ir(U-V)} = \mathbb{E}\, e^{irZ}, \quad r \in \mathbb{R}.$$

Conclude that $\sqrt{X}\, Y$ has the same distribution as $U - V$. (Was the attentive reader able to compute the density in Example 1.16? Can he do it now?)

2.31 *Laplace transforms and finiteness.* Recall that $0 \cdot x = 0$ for all $x \in \mathbb{R}$ and for $x = +\infty$. Thus, if $\hat{\mu}_r = \mathbb{E}\, e^{-rX}$ for some positive random variable X, then $\hat{\mu}_0 = 1$. Show that $r \mapsto \hat{\mu}_r$ is continuous and decreasing on $(0, \infty)$. Its continuity at 0 depends on whether X is almost surely finite: show that

$$\lim_{r \downarrow 0} \hat{\mu}_r = \mathbb{P}\{X < +\infty\}.$$

Hint: For $r > 0$, $e^{-rX} = e^{-rX} 1_{\{X<\infty\}} \nearrow 1_{\{X<\infty\}}$ as $r \downarrow 0$.

2.32 *Laplace transforms and moments.* Let $r \mapsto \hat{\mu}_r$ be the Laplace transform for a positive and almost surely finite random variable X. Use Fubini's theorem for the product measure $\mathbb{P} \times \text{Leb}$ to show that

$$\int_r^\infty dq\, \mathbb{E}\, Xe^{-qX} = \hat{\mu}_r, \qquad r \in \mathbb{R}_+.$$

This shows, when $\mathbb{E}X$ is finite, that the Laplace transform $\hat{\mu}$ is differentiable on $\mathbb{R}_+^* = (0, \infty)$, and

$$\frac{d}{dr}\hat{\mu}_r = -\mathbb{E}\, Xe^{-rX}, \qquad r \in \mathbb{R}_+^*;$$

in particular, then, the dominated convergence theorem yields

$$\lim_{r \downarrow 0} \frac{d}{dr}\hat{\mu}_r = -\mathbb{E}X.$$

A similar result holds for higher moments: if $\mathbb{E}X^n < \infty$,

$$\lim_{r \downarrow 0} \frac{d^n}{dr^n}\hat{\mu}_r = (-1)^n\, \mathbb{E}X^n.$$

2.33 *Characteristic functions and moments.* Let $\hat{\mu}$ be the characteristic function of a real-valued random variable X. Then, similar to the results of 2.32,

$$\lim_{r\to 0} \frac{d^n}{dr^n}\hat{\mu}_r = i^n \ \mathbb{E}X^n, \qquad n \in \mathbb{N},$$

provided that X^n be integrable, that is, provided that $\mathbb{E}\,|X|^n < \infty$. Generally, the equality above fails when $\mathbb{E}\,|X|^n = \infty$. However, for n even, if the limit on the left is finite, then the equality holds.

2.34 *Uniqueness of distributions and Laplace transforms.* Let X and Y be positive random variables. Show that the following are equivalent:

a) X and Y have the same distribution.
b) $\mathbb{E}\ e^{-rX} = \mathbb{E}\ e^{-rY}$ for every r in $\mathbb{R}_+$.
c) $\mathbb{E}\ f\circ X = \mathbb{E}\ f\circ Y$ for every f bounded continuous.
d) $\mathbb{E}\ f\circ X = \mathbb{E}\ f\circ Y$ for every f bounded Borel.
e) $\mathbb{E}\ f\circ X = \mathbb{E}\ f\circ Y$ for every f positive Borel.

Hint: Show that (a) $\Rightarrow$ (b) $\Rightarrow$ (c) $\Rightarrow$ (a) $\iff$ (e) $\iff$ (d). The difficult parts are (b) $\Rightarrow$ (c) and (c) $\Rightarrow$ (a). For (c) $\Rightarrow$ (a), start by showing that the indicator of an open interval is the limit of an increasing sequence of bounded continuous functions, and use the fact that open intervals form a p-system that generates the Borel σ-algebra on $\mathbb{R}$. For showing (b) $\Rightarrow$ (c), it is useful to recall the following consequence of the Stone-Weierstrass theorem: Let F be the collection of all functions f on $\bar{\mathbb{R}}_+$ having the form

$$f(x) = \sum_{i=1}^{n} c_i e^{-r_i x}$$

for some integer $n \geq 1$, constants $c_1, \ldots, c_n$ in $\mathbb{R}$, and constants $r_1, \ldots, r_n$ in $\mathbb{R}_+$. For every continuous function f on an interval $[a, b]$ of $\mathbb{R}_+$ there exists a sequence in F that converges to f uniformly on $[a, b]$.

2.35 *Uniqueness and characteristic functions.* Let X and Y be real-valued random variables. The statements (a)-(e) in the preceding exercise remain equivalent, except that (b) should be replaced with

b') $\mathbb{E}\ e^{irX} = \mathbb{E}\ e^{irY}$ for every r in $\mathbb{R}$.

2.36 *Random vectors.* Let $X = (X_1, \ldots, X_d)$ be a random variable taking values in $\mathbb{R}^d$, here $d \geq 1$ is an integer. The expected value of X is defined to be the vector

$$\mathbb{E}X = (\mathbb{E}X_1, \ldots, \mathbb{E}X_d).$$

The characteristic function of X is defined to be

$$\mathbb{E}\ e^{ir\cdot X}, \qquad r \in \mathbb{R}^d,$$

where $r \cdot x = r_1x_1 + \cdots + r_dx_d$, the inner product of r and x. When the components X_i are positive, Laplace transform of the distribution of X is defined similarly: $\mathbb{E}\ e^{-r\cdot X}$, $r \in \mathbb{R}^d_+$. As in the one-dimensional case, the characteristic function determines the distribution of X, and similarly for the Laplace transform. The equivalences in Exercises 2.34 and 2.35 remain true with the obvious modifications: in 2.34(b) and 2.35(b'), r should be in $\mathbb{R}^d_+$ and $\mathbb{R}^d$ respectively, and the functions alluded to should be defined on $\mathbb{R}^d_+$ and $\mathbb{R}^d$ respectively.

2.37 *Covariance.* Let X and Y be real-valued random variables with finite variances. Then, their covariance is defined to be

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

which is well-defined, is finite, and is bounded in absolute value by $\sqrt{\mathrm{Var}\ X}\ \sqrt{\mathrm{Var}\ Y}$; see Schwartz inequality in the next section. Show that $\mathrm{Var}(X+Y) = \mathrm{Var}\ X + \mathrm{Var}\ Y + 2\mathrm{Cov}(X,Y)$. If X and Y are independent, then $\mathrm{Cov}(X,Y) = 0$. The converse is generally false.

2.38 *Orthogonality.* Let X and Y be as in 2.37 above. They are said to be orthogonal, or *uncorrelated*, if $\mathbb{E}\ XY = \mathbb{E}X\ \mathbb{E}Y$. So, orthogonality is the same as having vanishing covariance. Show that, if $X_1, \ldots, X_n$ are pairwise orthogonal, that is, X_i and X_j are orthogonal for $i \neq j$, then

$$\mathrm{Var}(X_1 + \cdots + X_n) = \mathrm{Var}\ X_1 + \cdots + \mathrm{Var}\ X_n.$$

2.39 *Multi-dimensional Gaussian vectors.* Let X be a d-dimensional random vector; see 2.36 above. It is said to be *Gaussian* if $r \cdot X = r_1X_1 + \cdots + r_dX_d$ has a Gaussian distribution for every vector r in $\mathbb{R}^d$. It follows that the characteristic function of X has the form

$$\mathbb{E}\ e^{ir\cdot X} = e^{ia(r) - b(r)/2}, \quad r \in \mathbb{R}^d,$$

where $a(r) = \mathbb{E}\ r \cdot X$ and $b(r) = \mathrm{Var}\ r \cdot X$. Let

$$a = (a_1, \ldots, a_d) = (\mathbb{E}\ X_1, \ldots, \mathbb{E}\ X_d) = \mathbb{E}\ X,$$

and let $v = (v_{ij})$ be the $d \times d$ matrix of covariances $v_{ij} = \mathrm{Cov}(X_i, X_j)$. Note that the diagonal entries are variances.

a) Show that $a(r) = a \cdot r$ and $b(r) = r \cdot vr$ where vr is the vector obtained when v is multiplied by the column vector r. Conclude that the distribution of a Gaussian vector X is determined by its mean vector a and covariance matrix v.

b) Show that v is necessarily symmetric and positive definite, that is, $v_{ij} = v_{ji}$ for all i and j, and

$$r \cdot vr = \sum_{i=1}^{d} \sum_{j=1}^{d} r_i v_{ij} r_j \geq 0$$

for every r in $\mathbb{R}^d$.

2.40 *Independence.* Let X be a Gaussian random vector in $\mathbb{R}^d$ with mean vector a and covariance matrix v. Show that X_i and X_j are independent if and only if $v_{ij} = 0$. More generally, if I and J are disjoint subsets of $\{1, \ldots, d\}$, the random vectors $(X_i)_{i \in I}$ and $(X_j)_{j \in J}$ are independent if and only if $v_{ij} = 0$ for every pair (i, j) in $I \times J$. Show.

2.41 *Gaussian distribution.* Let X be a Gaussian vector in $\mathbb{R}^d$ with mean a and covariance matrix v. Then, its characteristic function is given by

$$\mathbb{E}\, e^{ir \cdot X} = e^{ia \cdot r - (r \cdot vr)/2}, \quad r \in \mathbb{R}^d.$$

If v is invertible, that is, if the rank of v is d, the distribution μ of X is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$, and the corresponding density function is

$$\frac{1}{\sqrt{\det(2\pi v)}} \exp[-\frac{1}{2}(x - a) \cdot v^{-1}(x - a)], \quad x \in \mathbb{R}^d,$$

where v^{-1} is the inverse of v and $\det m$ is the determinant of m; note that $\det(2\pi v) = (2\pi)^d \det v$.

If v is singular, that is, if the rank d' of v is less than d, then at least one entry of the vector X is a linear combination of the other entries. In that case, the distribution μ is no longer absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$. Instead, μ puts its mass on some hyperplane of dimension d' in $\mathbb{R}^d$.

2.42 *Continuation.* Let Z_1 and Z_2 be independent standard Gaussian variables (with means 0 and variances 1). Define a random vector X in $\mathbb{R}^3$ by letting $X = cZ$, where

$$X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}, \quad c = \begin{bmatrix} 1 & 2 \\ -1 & 3 \\ 4 & 1 \end{bmatrix}, \quad Z = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}$$

Each X_i is a linear combination of Z_1 and Z_2, therefore every linear combination of X_1, X_2, X_3 is also a linear combination of Z_1, Z_2. So, X is a 3-dimensional Gaussian random vector. Show that its covariance matrix is $v = cc^T$, where c^T is the transpose of c, that is, $v_{ij} = \sum_{k=1}^{2} c_{ik} c_{jk}$. Show that X_3 is a linear combination of X_1 and X_2. Show that Z_1 and Z_2 are linear combinations of X_1 and X_2; find the coefficients involved.

2.43 *Representation of Gaussian vectors.* Every Gaussian random vector X in $\mathbb{R}^d$ has the form

$$X = a + cZ,$$

where a is in $\mathbb{R}^d$, and c is a $d \times d'$ matrix, and Z is a random vector in $\mathbb{R}^{d'}$ whose coordinates are independent one-dimensional standard Gaussian variables. Then, X has mean a and covariance matrix $v = cc^T$.

3 L^p-spaces and Uniform Integrability

Let $(\Omega, \mathcal{H}, \mathbb{P})$ be a probability space. Let X be a real-valued random variable. For p in $[1, \infty)$, define

$$\|X\|_p = (\mathbb{E}\,|X|^p)^{1/p}, \tag{3.1}$$

and for $p = \infty$ let

$$\|X\|_\infty = \inf\{b \in \mathbb{R}_+ :\ |X| \le b \text{ almost surely}\}. \tag{3.2}$$

It is easy to see that

$$\|X\|_p = 0 \quad \Rightarrow \quad X = 0 \text{ almost surely}, \tag{3.3}$$

$$\|cX\|_p = c\ \|X\|_p\,, \qquad c \ge 0; \tag{3.4}$$

and it will follow from Theorem 3.6a below with $Y = 1$ that

$$0 \le \|X\|_p \le \|X\|_q \le +\infty \qquad \text{if } 1 \le p \le q \le +\infty. \tag{3.5}$$

For each p in $[1, \infty]$, let L^p denote the collection of all real-valued random variables X with $\|X\|_p < \infty$. For p in $[1, \infty)$, X is in L^p if and only if $|X|^p$ is integrable; and X is in L^∞ if and only if X is almost surely bounded. For X in L^p, the number $\|X\|_p$ is called the *L^p-norm* of X; in particular, $\|X\|_\infty$ is called the *essential supremum* of X. Indeed, the properties 3.4 and 3.5 together with Minkowski's inequality below imply that each L^p is a normed vector space provided that we identify X and Y in L^p as one random variable if $X = Y$ almost surely.

Inequalities

The following theorem summarizes the various connections. Its proof will be put after a lemma of independent interest.

3.6 Theorem. *a)* Hölder's inequality*: For p, q, r in $[1, \infty)$ with $\frac{1}{p} + \frac{1}{q} = \frac{1}{r}$,*

$$\|XY\|_r \le \|X\|_p\, \|Y\|_q.$$

In particular, Schwartz's inequality *holds:* $\|XY\|_1 \le \|X\|_2\, \|Y\|_2$.

b) Minkowski's inequality*: For p in $[1, \infty]$,*

$$\|X + Y\|_p \le \|X\|_p + \|Y\|_p$$

3.7 Lemma. Jensen's inequality. *Let D be a convex domain in $\mathbb{R}^d$. Let $f : D \mapsto \mathbb{R}$ be continuous and concave. Suppose that $X_1, \ldots, X_d$ are integrable random variables and that the vector $(X_1, \ldots, X_d)$ belongs to D almost surely. Then,*

$$\mathbb{E}\, f(X_1, \ldots, X_d) \le f(\mathbb{E}X_1, \ldots \mathbb{E}X_d).$$

Proof. Since D is convex and $X = (X_1, \dots, X_d)$ is in D almost surely, the vector $a = \mathbb{E}X = (\mathbb{E}X_1, \dots \mathbb{E}X_d)$ belongs to D. Let $c_1, \dots, c_d$ be the direction cosines of a hyperplane in $\mathbb{R}^{d+1}$ lying above the surface f and passing through the point $(a, f(a))$ in $\mathbb{R}^d \times \mathbb{R}$. Then,

$$f(x) \le f(a) + \sum_1^d (x_i - a_i)\, c_i \qquad x \in D.$$

Replacing x with X and taking expectations yields the desired result. □

In preparation for the proof of Theorem 3.6, we leave it as an exercise to show that, for b in $(0, 1]$,

3.8 $$f(u, v) = u^b\, v^{1-b}, \qquad g(u, v) = (u^b + v^b)^{\frac{1}{b}}$$

define functions that are continuous and concave on $\mathbb{R}^2_+$. Thus, by the preceding lemma,

3.9 $$\mathbb{E}\, U^b V^{1-b} \le (\mathbb{E}U)^b\, (\mathbb{E}V)^{1-b}, \qquad \mathbb{E}\, (U^b + V^b)^{\frac{1}{b}} \le [(\mathbb{E}U)^b + (\mathbb{E}V)^b]^{\frac{1}{b}},$$

provided that U and V be positive integrable random variables.

Proof of Theorem 3.6

a) Hölder's inequality. Assume that $\|X\|_p$ and $\|Y\|_q$ are finite; otherwise, there is nothing to prove. When $p = \infty$, we have $|XY| \le \|X\|_p\, |Y|$ almost surely, and hence the inequality is immediate; similarly for $q = \infty$. Assuming that p and q are both finite, the inequality desired follows from the first inequality in 3.9 with $b = \frac{r}{p}$, $U = |X|^p$, $V = |Y|^q$.

b) Minkowski's inequality. Again, assume that $\|X\|_p$ and $\|Y\|_p$ are finite. If $p = \infty$, the inequality is immediate from the definition 3.2 (and in fact, becomes an equality). For p in $[1, \infty)$, the inequality follows from the second inequality in 3.9 with $b = \frac{1}{p}$, $U = |X|^p$, $V = |Y|^p$. □

Uniform integrability

This concept plays an important role in martingale theory and in the convergence of sequences in the space L^1. We start by illustrating the issue involved in the simplest setting.

3.10 Lemma. *Let X be a real-valued random variable. Then, X is integrable if and only if*

3.11 $$\lim_{b \to \infty} \mathbb{E}\, |X|\, 1_{\{|X| > b\}} = 0.$$

Proof. Let Z_b denote the variable inside the expectation in 3.11. Note that it is dominated by $|X|$ and goes to 0 as $b \to \infty$. Thus, if X is integrable, the dominated convergence yields that $\lim_{b\to\infty} \mathbb{E}\ Z_b = 0$, which is exactly 3.11. Conversely, if 3.11 holds, then we can choose b large enough to have $\mathbb{E}\ Z_b \leq 1$, and the inequality $|X| \leq b + Z_b$ shows that $\mathbb{E}\ |X| \leq b + 1 < \infty$. □

For a collection of random variables X, the uniform integrability of the collection has to do with the possibility of taking the limit in 3.11 uniformly in X:

3.12 Definition. *A collection $\mathcal{K}$ of real-valued random variables is said to be* uniformly integrable *if*

$$k(b) = \sup_{X\in\mathcal{K}} \mathbb{E}\ |X|\ 1_{\{|X|>b\}}$$

goes to 0 as $b \to \infty$.

3.13 Remarks. a) If $\mathcal{K}$ is finite and each X in it is integrable, then $\mathcal{K}$ is uniformly integrable. For, then, the limit over b, of $k(b)$, can be passed inside the supremum, and Lemma 3.10 does the rest.

b) If $\mathcal{K}$ is dominated by an integrable random variable Z, then it is uniformly integrable. Because, then, $|X| \leq Z$ for every X in $\mathcal{K}$, which yields $k(b) \leq \mathbb{E}\ Z\ 1_{\{Z>b\}}$, and that last expectation goes to 0 by Lemma 3.10 applied to Z.

c) Uniform integrability implies L^1-boundedness, that is, if $\mathcal{K}$ is uniformly integrable then $\mathcal{K} \subset L^1$ and

$$k(0) = \sup_{\mathcal{K}} \mathbb{E}\ |X| < \infty.$$

To see this, note that $\mathbb{E}\ |X| \leq b + k(b)$ for all X and use the uniform integrability of $\mathcal{K}$ to choose a finite number b such that $k(b) \leq 1$.

d) But L^1-boundedness is insufficient for uniform integrability. Here is a sequence $\mathcal{K} = \{X_n : n \geq 1\}$ that is L^1-bounded but not uniformly integrable. Suppose that $\Omega = (0, 1)$ with its Borel σ-algebra for events and the Lebesgue measure as $\mathbb{P}$. Let $X_n(\omega)$ be equal to n if $\omega \leq 1/n$ and to 0 otherwise. Then, $\mathbb{E}\ X_n = 1$ for all n, that is, $\mathcal{K}$ is L^1-bounded. But $k(b) = 1$ for all b, since $\mathbb{E}\ X_n\ 1_{\{X_n>b\}} = \mathbb{E}\ X_n = 1$ for $n > b$.

e) However, if $\mathcal{K}$ is L^p-bounded for some $p > 1$ then it is uniformly integrable. This will be shown below: see Proposition 3.17 and take $f(x) = x^p$.

The following ε-δ characterization is the main result on uniform integrability: over every small set, the integrals of the X are uniformly small.

3.14 Theorem. *The collection $\mathcal{K}$ is uniformly integrable if and only if it is L^1-bounded and for every $\varepsilon > 0$ there is $\delta > 0$ such that, for every event H,*

$$3.15 \qquad \mathbb{P}(H) \leq \delta \Rightarrow \sup_{X\in\mathcal{K}} \mathbb{E}\ |X|\ 1_H \leq \varepsilon.$$

Proof. We assume that all X are positive; this amounts to working with $|X|$ throughout. Since $X\ 1_H \le b\ 1_H + X\ 1_{\{X>b\}}$ for every event H and every b in $\mathbb{R}_+$,

$$\sup_{X\in\mathcal{K}} \mathbb{E}\ X\ 1_H \le b\mathbb{P}(H) + k(b), \quad b \in \mathbb{R}_+. \tag{3.16}$$

Suppose that $\mathcal{K}$ is uniformly integrable. Then, it is L^1-bounded by Remark 3.13c. Also, since $k(b) \to 0$, for every $\varepsilon > 0$ there is $b < \infty$ such that $k(b) \le \varepsilon/2$, and setting $\delta = \varepsilon/2b$ we see that 3.15 holds in view of 3.16.

Conversely, suppose that $\mathcal{K}$ is L^1-bounded and that for every $\varepsilon > 0$ there is $\delta > 0$ such that 3.15 holds for all events H. Then, Markov's inequality 2.17 yields

$$\sup_{X\in\mathcal{K}} \mathbb{P}\{X > b\} \le \frac{1}{b} \sup_{X\in\mathcal{K}} \mathbb{E}\ X = \frac{1}{b}k(0),$$

which shows the existence of b such that $\mathbb{P}\{X > b\} \le \delta$ for all X, and, then, for that b we have $k(b) \le \varepsilon$ in view of 3.15 used with $H = \{X > b\}$. In other words, for every $\varepsilon > 0$ there is $b < \infty$ such that $k(b) \le \varepsilon$, which is the definition of uniform integrability. □

The following proposition is very useful for showing uniform integrability. In particular, as remarked earlier, it shows that L^p-boundedness for some $p > 1$ implies uniform integrability.

3.17 Proposition. *Suppose that there is a positive Borel function f on $\mathbb{R}_+$ such that $\lim_{x\to\infty} f(x)/x = \infty$ and*

$$\sup_{X\in\mathcal{K}} \mathbb{E}\ f\circ|X| < \infty. \tag{3.18}$$

Then, $\mathcal{K}$ is uniformly integrable.

Proof. We may and do assume that all X are positive. Also, by replacing f with $f \vee 1$ if necessary, we assume that $f \ge 1$ in addition to satisfying the stated conditions. Let $g(x) = x/f(x)$ and note that

$$X\ 1_{\{X>b\}} = f\circ X\ g\circ X\ 1_{\{X>b\}} \le f\circ X \sup_{x>b} g(x).$$

This shows that, with c denoting the supremum in 3.18,

$$k(b) \le c \sup_{x>b} g(x),$$

and the right side goes to 0 as $b \to \infty$ since $g(x) \to 0$ as $x \to +\infty$. □

We supplement the preceding proposition by a converse and give another characterization.

3.19 THEOREM. *The following are equivalent:*

a) $\mathcal{K}$ *is uniformly integrable.*
b) $h(b) = \sup_{\mathcal{K}} \int_b^\infty dy\ \mathbb{P}\{|X| > y\} \to 0$ *as* $b \to \infty$.
c) $\sup_{\mathcal{K}} \mathbb{E}\ f \circ |X| < \infty$ *for some increasing convex function* f *on* $\mathbb{R}_+$ *with* $\lim_{x\to\infty} f(x)/x = +\infty$.

Proof. The preceding proposition shows that (c) ⇒ (a). We now show that (a) ⇒ (b) ⇒ (c), again assuming, as we may, that all the X in $\mathcal{K}$ are positive.

Assume (a). For every X in $\mathcal{K}$,

$$\begin{aligned}\mathbb{E}\ X\ 1_{\{X>b\}} &= \int_0^\infty dy\ \mathbb{P}\{X\ 1_{\{X>b\}} > y\} \\ &= \int_0^\infty dy\ \mathbb{P}\{X > b \vee y\}\ \ge\ \int_b^\infty dy\ \mathbb{P}\{X > y\}.\end{aligned}$$

Thus, $k(b) \ge h(b)$ for every b, and the uniform integrability of $\mathcal{K}$ means that $k(b) \to 0$ as $b \to \infty$. Hence, (a) ⇒ (b).

Assume (b). Since $h(b) \to 0$ as $b \to \infty$, we can pick $0 = b_0 < b_1 < b_2 < \cdots$ increasing to $+\infty$ such that

$$h(b_n) \le h(0)/2^n, \quad n \in \mathbb{N};$$

note that $h(0)$ is finite since $h(0) \le b + h(b)$ and $h(b)$ can be made as small as desired. Define

$$g(x) = \sum_{n=0}^\infty 1_{[b_n,\infty)}(x), \quad f(x) = \int_0^x dy\ g(y), \quad x \in \mathbb{R}_+.$$

Note that $g \ge 1$ and is increasing toward $+\infty$, which implies that f is increasing and convex and $\lim_{x\to\infty} f(x)/x = +\infty$. Now,

$$\begin{aligned}\mathbb{E}\ f\circ X &= \mathbb{E}\ \int_0^X dy\ g(y) \\ &= \sum_{n=0}^\infty\ \mathbb{E}\ \int_{b_n}^\infty dy\ 1_{\{X>y\}}\ \le\ \sum_{n=0}^\infty h(b_n)\ \le\ 2h(0)\ <\ \infty.\end{aligned}$$

This being true for all X in $\mathcal{K}$, we see that (b) ⇒ (c). □

Exercises and complements

3.20 *Concavity.* Show that the functions f and g defined by 3.8 are continuous and concave. Hint: Note that $f(cu, cv) = c\,f(u, v)$ for every $c > 0$; conclude that it is sufficient to show that $x \mapsto f(x, 1-x)$ from $[0, 1]$ into $\mathbb{R}_+$ is continuous and concave; and show the latter by noting that the second derivative is negative. Similarly for g.

3.21 *Continuity of the norms.* Fix a random variable X. Define $f(p) = \|X\|_p$ for p in $[1, \infty]$. Show that the function f is continuous except possibly at one point p_0, where p_0 is such that

$$f(p) < \infty \text{ for } p < p_0\,, \qquad f(p) = +\infty \text{ for } p > p_0\,,$$

and f is left-continuous at p_0.

3.22 *Integrals over small sets.* Let X be positive and integrable. Let (H_n) be a sequence of events. If $\mathbb{P}(H_n) \to 0$, then $\mathbb{E}X1_{H_n} \to 0$. Show.

3.23 *Uniform integrability.* Let (X_i) and (Y_i) be uniformly integrable. Show that, then,

a) $(X_i \vee Y_i)$ is uniformly integrable,
b) $(X_i + Y_i)$ is uniformly integrable.

3.24 *Comparisons.* If $|X_i| \leq |Y_i|$ for each i, and (Y_i) is uniformly integrable, then so is (X_i). Show.

4 Information and Determinability

This section is on σ-algebras generated by random variables and measurability with respect to them. Also, we shall argue that such a σ-algebra should be thought as a body of information, and measurability with respect to it should be equated to being determined by that information. Throughout, $(\Omega, \mathcal{H}, \mathbb{P})$ is a probability space.

Sigma-algebras generated by random variables

Let X be a random variable taking values in some measurable space $(E, \mathcal{E})$. Then,

4.1 $$\sigma X = X^{-1}\mathcal{E} = \{X^{-1}A : \quad A \in \mathcal{E}\}$$

is a σ-algebra (and is a subset of $\mathcal{H}$ by the definition of random variables). It is called the *σ-algebra generated by X*, and the notation σX is preferred over the others. Clearly, σX is the smallest σ-algebra $\mathcal{G}$ on Ω such that X is measurable with respect to $\mathcal{G}$ and $\mathcal{E}$; see Exercise I.2.20.

Let T be an arbitrary index set, countable or uncountable. For each t in T let X_t be a random variable taking values in some measurable space $(E_t, \mathcal{E}_t)$. Then,

4.2 $$\sigma\{X_t : \quad t \in T\} = \bigvee_{t \in T} \sigma X_t$$

denotes the σ-algebra on Ω generated by the union of the σ-algebras σX_t, $t \in T$; see Exercise I.1.18. It is called the *σ-algebra generated* by the collection $\{X_t : \ t \in T\}$. It is the smallest σ-algebra $\mathcal{G}$ on Ω such that, for every t in T, the random variable X_t is measurable with respect to $\mathcal{G}$ and $\mathcal{E}_t$; obviously, $\mathcal{G} \subset \mathcal{H}$.

In view of Proposition I.6.27, we may regard the collection $\{X_t : t \in T\}$ as one random variable X taking values in the product space $(E, \mathcal{E}) = \otimes_{t\in T}(E_t, \mathcal{E}_t)$ by defining $X(\omega)$ to be the point $(X_t(\omega))_{t\in T}$ in the "function" space E for each ω. Conversely, if X is a random variable taking values in the product space $(E, \mathcal{E})$, we denote by $X_t(\omega)$ the value of the function $X(\omega)$ at the point t in T; the resulting mapping $\omega \mapsto X_t(\omega)$ is a random variable with values in $(E_t, \mathcal{E}_t)$ and is called the *t-coordinate* of X. It will be convenient to write $X = (X_t)_{t\in T}$ and consider X both as the E-valued random variable and as the collection of random variables X_t, $t \in T$. This causes no ambiguity for σX:

4.3 PROPOSITION. *If $X = (X_t)_{t\in T}$, then $\sigma X = \sigma\{X_t : t \in T\}$.*

Proof. Proof is immediate from that of Proposition I.6.27. Let $\mathcal{H}$ there be σX to conclude that $\sigma X \supset \sigma\{X_t : t \in T\}$, and then let $\mathcal{H}$ be $\sigma\{X_t : t \in T\}$ to conclude that $\sigma\{X_t : t \in T\} \supset \sigma X$. □

Measurability

The following theorem is to characterize the σ-algebra σX. It shows that a random variable is σX-measurable if and only if it is a deterministic measurable function of X. In other words, with the usual identification of a σ-algebra with the collection of all numerical mappings that are measurable relative to it, the collection σX of random variables is exactly the set of all measurable functions of X.

4.4 THEOREM. *Let X be a random variable taking values in some measurable space $(E, \mathcal{E})$. A mapping $V : \Omega \to \bar{\mathbb{R}}$ belongs to σX if and only if*

$$V = f \circ X$$

for some deterministic function f in $\mathcal{E}$.

Proof. Sufficiency. Since X is measurable with respect to σX and $\mathcal{E}$, and since measurable functions of measurable functions are measurable, every V having the form $f \circ X$ for some f in $\mathcal{E}$ is σX-measurable.

Necessity. Let $\mathcal{M}$ be the collection of all V having the form $V = f \circ X$ for some f in $\mathcal{E}$. We shall use the monotone class theorem I.2.19 to show that $\mathcal{M} \supset \sigma X$, which is the desired result. We start by showing that $\mathcal{M}$ is a monotone class of functions on Ω.

i) $1 \in \mathcal{M}$ since $1 = f \circ X$ with $f(x) = 1$ for all x in E.

ii) Let U and V be bounded and in $\mathcal{M}$, and let a and b be in $\mathbb{R}$. Then, $U = f \circ X$ and $V = g \circ X$ for some f and g in $\mathcal{E}$, and thus, $aU + bV = h \circ X$ with $h = af + bg$. Since $h \in \mathcal{E}$, it follows that $aU + bV \in \mathcal{M}$.

iii) Let $(V_n) \subset \mathcal{M}_+$ and $V_n \nearrow V$. For each n, there is f_n in $\mathcal{E}$ such that $V_n = f_n \circ X$. Then, $f = \sup f_n$ belongs to $\mathcal{E}$ and since $V_n \nearrow V$,

$$V(\omega) = \sup_n V_n(\omega) = \sup_n f_n(X(\omega)) = f(X(\omega)) , \qquad \omega \in \Omega,$$

which shows that $V \in \mathcal{M}$.

Furthermore, $\mathcal{M}$ includes every indicator variable in σX: if $H \subset \Omega$ is in σX, then $H = X^{-1}A$ for some set A in $\mathcal{E}$, and $1_H = 1_A \circ X \in \mathcal{M}$. Therefore, by the monotone class theorem, $\mathcal{M}$ contains all positive random variables in σX.

Finally, let V in σX be arbitary. Then, $V^+ \in \sigma X$ and is positive, and hence, $V^+ = g \circ X$ for some g in $\mathcal{E}$; similarly, $V^- = h \circ X$ for some h in $\mathcal{E}$. Thus, $V = V^+ - V^- = f \circ X$, where

$$f(x) = \begin{cases} g(x) - h(x) & \text{if } g(x) \wedge h(x) = 0, \\ 0 & \text{otherwise.} \end{cases}$$

This completes the proof since $f \in \mathcal{E}$. □

4.5 Corollary. *For each n in $\mathbb{N}^*$, let X_n be a random variable taking values in some measurable space $(E_n, \mathcal{E}_n)$. A mapping $V : \Omega \mapsto \bar{\mathbb{R}}$ belongs to $\sigma\{X_n : n \in \mathbb{N}^*\}$ if and only if*

$$V = f(X_1, X_2, \ldots)$$

for some f in $\otimes_n \mathcal{E}_n$.

Proof. Proof is immediate from the preceding theorem upon putting $X = (X_1, X_2, \ldots)$ and using Proposition 4.3. □

The preceding corollary can be generalized to uncountable collections $\{X_t : t \in T\}$ by using the same device of regarding the collection as one random variable. In fact, there is a certain amount of simplification, reflecting the fact that uncountable products of σ-algebras $\mathcal{E}_t$, $t \in T$, are in fact generated by the finite-dimensional rectangles.

4.6 Proposition. *Let T be arbitrary. For each t in T, let X_t be a random variable taking values in some measurable space $(E_t, \mathcal{E}_t)$. Then, $V : \Omega \mapsto \bar{\mathbb{R}}$ belongs to $\sigma\{X_t : t \in T\}$ if and only if there exists a sequence (t_n) in T and a function f in $\otimes_n \mathcal{E}_{t_n}$ such that*

4.7 $$V = f(X_{t_1}, X_{t_2}, \ldots).$$

Proof. Sufficiency of the condition is trivial: if V has the form 4.7, then $V \in \sigma\{X_{t_n} : n \geq 1\} = \hat{\mathcal{G}}$ by the corollary above, and $\hat{\mathcal{G}} \subset \mathcal{G} = \sigma\{X_t : t \in T\}$ obviously.

To show the necessity, we use the monotone class theorem I.2.19 together with Proposition 4.3. To that end, let $\mathcal{M}$ be the collection of all V having the form 4.7 for some sequence (t_n) in T and some f in $\otimes\mathcal{E}_{t_n}$. It is easy to check that $\mathcal{M}$ is a monotone class. We shall show that $\mathcal{M}$ includes the indicators of a p-system $\mathcal{G}_0$ that generates $\mathcal{G}$. Then, by the monotone class theorem, $\mathcal{M}$ includes all positive V in $\mathcal{G}$ and, therefore, all V in $\mathcal{G}$ since $V = V^+ - V^-$ is obviously in $\mathcal{M}$ if V^+ and V^- are in $\mathcal{M}$. Hence, $\mathcal{M} \supset \mathcal{G}$ as desired.

By Proposition 4.3, $\mathcal{G} = \sigma X$ where $X = (X_t)_{t\in T}$ takes values in $(E, \mathcal{E}) = \bigotimes(E_t, \mathcal{E}_t)$. Recall that $\mathcal{E}$ is generated by the p-system of all finite-dimensional measurable rectangles. Therefore, the inverse images $X^{-1}A$ of those rectangles A form a p-system $\mathcal{G}_0$ that generates $\mathcal{G}$. Thus, to complete the proof, it is sufficient to show that the indicator of $X^{-1}A = \{X \in A\}$ belongs to $\mathcal{M}$ for every such rectangle A.

Let A be such a rectangle, that is, $A = \bigtimes_t A_t$ with $A_t = E_t$ for all t outside a finite subset S of T and $A_t \in \mathcal{E}_t$ for every t in S (and therefore for all t in T). Then,

$$1_{\{X\in A\}} = 1_A \circ X = \prod_{t\in S} 1_{A_t} \circ X_t,$$

which has the form 4.7, that is, belongs to $\mathcal{M}$. □

Heuristics

Our aim is to use the foregoing to argue that a σ-algebra on Ω is the mathematically precise equivalent of the everyday term "information". And, random quantities that are determined by that information are precisely the random variables that are measurable with respect to that σ-algebra.

To fix the ideas, consider a random experiment that consists of a sequence of trials, at each of which there are five possible results labeled a, b, c, d, e. Each possible outcome of this experiment can be represented by a sequence $\omega = (\omega_1, \omega_2, \ldots)$ where $\omega_n \in E = \{a, \ldots, e\}$ for each n. The sample space Ω, then, consists of all such sequences ω. We define $X_1, X_2, \ldots$ to be the coordinate variables, that is $X_n(\omega) = \omega_n$ for every n and outcome ω. We let $\mathcal{H}$ be the σ-algebra generated by $\{X_n : n \in \mathbb{N}^*\}$. The probability $\mathbb{P}$ is unimportant for our current purposes and we leave it unspecified.

Consider the information we shall have about this experiment at the end of the third trial. At that time, whatever the possible outcome ω may be, we shall know $X_1(\omega)$, $X_2(\omega)$, $X_3(\omega)$, and nothing more. In other words, the information we shall have will specify the results $\omega_1, \omega_2, \omega_3$ but nothing more. Thus, the information we shall have will determine the values $V(\omega)$, for every possible ω, provided that the dependence of $V(\omega)$ on ω is through $\omega_1, \omega_2, \omega_3$, that is, provided that $V = f(X_1, X_2, X_3)$ for some deterministic function f on $E\times E\times E$. Based on these arguments, we equate "the information available at the end of third trial" to the σ-algebra $\mathcal{G}$ consisting of all such numerical random variables whose values are determined by that body of information.

In this case, the information $\mathcal{G}$ is generated by $\{X_1, X_2, X_3\}$ in the sense that knowing X_1, X_2, X_3 is equivalent to knowing the information $\mathcal{G}$.

Going back to an arbitrary probability space $(\Omega, \mathcal{H}, \mathbb{P})$ and a sub-σ-algebra $\mathcal{G}$ of $\mathcal{H}$, we may heuristically equate $\mathcal{G}$ to the information available to someone who is able to tell the value $V(\omega)$ for every possible ω and every random variable V that is $\mathcal{G}$-measurable. Incidentally, this gives a mathematical definition for the imprecise everyday term "information".

Often, there are simpler ways of characterizing the information $\mathcal{G}$. If there is a random variable X such that the knowledge of its value is sufficient to determine the values of all the V in $\mathcal{G}$, then we say that X generates the information $\mathcal{G}$ and write $\mathcal{G} = \sigma X$. This is the heuristic content of the definition of σX.

Of course, embedded in the heuristics is the basic theorem of this section, Theorem 4.4, which now becomes obvious: if the information $\mathcal{G}$ consists of the knowledge of X, then $\mathcal{G}$ determines exactly those variables V that are deterministic functions of X. Another result that becomes obvious is Proposition 4.3: in the setting of it, since knowing X is the same as knowing X_t for all t in T, the information generated by X is the same as the information generated by $X_t,\ t \in T$.

Filtrations

Continuing with the heuristics, suppose that we are interested in a random experiment taking place over an infinite expanse of time. Let $T = \mathbb{R}_+$ or $T = \mathbb{N}$ be the time set. For each time t, let $\mathcal{F}_t$ be the information gathered during $[0, t]$ by an observer of the experiment. For $s < t$, we must have $\mathcal{F}_s \subset \mathcal{F}_t$. The family $\mathcal{F} = \{\mathcal{F}_t :\ t \in T\}$, then, depicts the flow of information as the experiment progresses over time. The following definition formalizes this concept.

4.8 Definition. *Let T be a subset of $\mathbb{R}$. For each t in T, let $\mathcal{F}_t$ be a sub-σ-algebra of $\mathcal{H}$. The family $\mathcal{F} = \{\mathcal{F}_t :\ t \in T\}$ is called a* filtration *provided that $\mathcal{F}_s \subset \mathcal{F}_t$ for $s < t$.*

In other words, a filtration is an increasing family of sub-σ-algebras of $\mathcal{H}$. The simplest examples are the filtrations generated by stochastic processes: If $X = \{X_t :\ t \in T\}$ is a stochastic process, then putting $\mathcal{F}_t = \sigma\{X_s :\ s \le t,\ s \in T\}$ yields a filtration $\mathcal{F} = \{\mathcal{F}_t :\ t \in T\}$. The reader is invited to ponder the meaning of the next proposition for such a filtration. Of course, the aim is to approximate eternal variables by random variables that become known in finite time.

4.9 Proposition. *Let $\mathcal{F} = \{\mathcal{F}_n :\ n \in \mathbb{N}\}$ be a filtration and put $\mathcal{F}_\infty = \bigvee_{n\in\mathbb{N}} \mathcal{F}_n$. For each bounded random variable V in $\mathcal{F}_\infty$ there are bounded variables V_n in $\mathcal{F}_n$, $n \in \mathbb{N}$, such that*

$$\lim_n \mathbb{E}\,|V_n - V| \;=\; 0.$$

REMARK. Note that $\mathbb{E}|V_n - V| = \|V_n - V\|_1$ in the notation of section 3; thus, the approximation here is in the sense of L^1-space. Also, we may add to the conclusion that $\mathbb{E}V_n \to \mathbb{E}V$; this follows from the observation that $|\mathbb{E}V_n - \mathbb{E}V| \le \mathbb{E}|V_n - V|$.

Proof. Let $\mathcal{C} = \bigcup_n \mathcal{F}_n$. By definition, $\mathcal{F}_\infty = \sigma\mathcal{C}$. Obviously $\mathcal{C}$ is a p-system. To complete the proof via the monotone class theorem, we start by letting $\mathcal{M}_b$ be the collection of all bounded variables in $\mathcal{F}_\infty$ having the approximation property described. It is easy to see that $\mathcal{M}_b$ includes constants and is a vector space over $\mathbb{R}$ and includes the indicators of events in $\mathcal{C}$. Thus, $\mathcal{M}_b$ will include all bounded V in $\mathcal{F}_\infty$ once we check the remaining monotonicity condition.

Let $(U_k) \subset \mathcal{M}_b$ be positive and increasing to a bounded variable V in $\mathcal{F}_\infty$. Then, for each $k \ge 1$ there are $U_{k,n}$ in $\mathcal{F}_n$, $n \in \mathbb{N}$, such that $\mathbb{E}|U_{k,n} - U_k| \to 0$ as $n \to \infty$. Put $n_0 = 0$, and for each $k \ge 1$ choose $n_k > n_{k-1}$ such that $\hat{U}_k = U_{k,n_k}$ satisfies

$$\mathbb{E}|\hat{U}_k - U_k| < \frac{1}{k}.$$

Moreover, since (U_k) is bounded and converges to V, the bounded convergence implies that $\mathbb{E}|U_k - V| \to 0$. Hence,

$$\mathbb{E}\,|\hat{U}_k - V| \;\le\; \mathbb{E}\,|\hat{U}_k - U_k| \;+\; \mathbb{E}\,|U_k - V| \;\to\; 0 \tag{4.10}$$

as $k \to \infty$. With $n_0 = 0$ choose $V_0 = 0$ and put $V_n = \hat{U}_k$ for all integers n in $(n_k, n_{k+1}]$; then, $V_n \in \mathcal{F}_{n_k} \subset \mathcal{F}_n$, and $\mathbb{E}|V_n - V| \to 0$ as $n \to \infty$ in view of 4.10. This is what we need to show that $V \in \mathcal{M}_b$. □

In the preceding proposition, the V_n are shown to exist but are unspecified. A very specific version will appear later employing totally new tools; see the martingale convergence theorems of Chapter V and, in particular, Corollary V.3.30 there.

Exercises and complements

4.11 *p-systems for σX.* Let T be an arbitrary index set. Let $X = (X_t)_{t\in T}$, where X_t takes values in $(E_t, \mathcal{E}_t)$ for each t in T. For each t, let $\mathcal{C}_t$ be a p-system that generates $\mathcal{E}_t$. Let $\mathcal{G}_0$ be the collection of all $G \subset \Omega$ having the form

$$G = \bigcap_{t\in S} \{X_t \in A_t\}$$

for some finite $S \subset T$ and A_t in $\mathcal{C}_t$ for every t in S. Show that $\mathcal{G}_0$ is a p-system that generates $\mathcal{G} = \sigma X$.

4.12 *Monotone class theorem.* This is a generalization of the monotone class theorem I.2.19. We keep the setting and notations of the preceding exercise.

Let $\mathcal{M}$ be a monotone class of mappings from Ω into $\bar{\mathbb{R}}$. Suppose that $\mathcal{M}$ includes every $V : \Omega \mapsto [0,1]$ having the form

$$V = \prod_{t \in S} 1_{A_t} \circ X_t \,, \qquad S \text{ finite, } A_t \in \mathcal{C}_t \text{ for every } t \text{ in } S.$$

Then, every positive V in σX belongs to $\mathcal{M}$. Prove.

4.13 *Special case.* In the setting of the exercises above, suppose $E_t = \mathbb{R}$ and $\mathcal{E}_t = \mathcal{B}_{\mathbb{R}}$ for all t. Let $\mathcal{M}$ be a monotone class of mappings from Ω into $\bar{\mathbb{R}}$. Suppose that $\mathcal{M}$ includes every V of the form

$$V = f_1 \circ X_{t_1} \cdots f_n \circ X_{t_n}$$

with $n \geq 1$ and $t_1, \ldots, t_n$ in T and $f_1, \ldots, f_n$ bounded continuous functions from $\mathbb{R}$ into $\mathbb{R}$. Then, $\mathcal{M}$ contains all positive V in σX. Prove. Hint: Start by showing that, if A is an open interval of $\mathbb{R}$, then 1_A is the limit of an increasing sequence of bounded continuous functions.

4.14 *Determinability.* If X and Y are random variables taking values in $(E, \mathcal{E})$ and $(D, \mathcal{D})$, then we say that X *determines* Y if $Y = f \circ X$ for some $f : E \mapsto D$ measurable with respect to $\mathcal{E}$ and $\mathcal{D}$. Then, $\sigma X \supset \sigma Y$ obviously. Heuristically, X determines Y if knowing $X(\omega)$ is sufficient for knowing $Y(\omega)$, this being true for every possibility ω. To illustrate the notion in a simple setting, let T be a positive random variable and define a stochastic process $X = (X_t)_{t \in \mathbb{R}_+}$ by setting, for each ω

$$X_t(\omega) = \begin{cases} 0 & \text{if } t < T(\omega), \\ 1 & \text{if } t \geq T(\omega). \end{cases}$$

Show that X and T determine each other. If T represents the time of failure for a device, then X is the process that indicates whether the device has failed or not. That X and T determine each other is intuitively obvious, but the measurability issues cannot be ignored altogether.

4.15 *Warning.* A slight change in the preceding exercise shows that one must guard against raw intuition. Let T have a distribution that is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}_+$; in fact, all we need is that $\mathbb{P}\{T = t\} = 0$ for every t in $\mathbb{R}_+$. Define

$$X_t(\omega) = \begin{cases} 1 & \text{if } t = T(\omega) \\ 0 & \text{otherwise.} \end{cases}$$

Show that, for each t in $\mathbb{R}_+$, the random variable X_t is determined by T. But, contrary to raw intuition, T is not determined by $X = (X_t)_{t \in \mathbb{R}_+}$. Show this by following the steps below:

a) For each t, we have $X_t = 0$ almost surely. Therefore, for every sequence (t_n) in $\mathbb{R}_+$, $X_{t_1} = X_{t_2} = \ldots = 0$ almost surely.

b) If $V \in \sigma X$, then $V = c$ almost surely for some constant c. It follows that T is not in σX.

4.16 *Arrival processes.* Let $T = (T_1, T_2, \ldots)$ be an increasing sequence of $\mathbb{R}_+$-valued variables. Define a stochastic process $X = (X_t)_{t\in\mathbb{R}_+}$ with state space $\mathbb{N}$ by

$$X_t = \sum_{n=1}^{\infty} 1_{(0,t]} \circ T_n \ , \qquad t \in \mathbb{R}_+.$$

Show that X and T determine each other. If T_n represents the n-th arrival time at a store, then X_t is the number of customers who arrived during $(0, t]$. So, X and T are the same phenomena viewed from different angles.

5 Independence

This section is about independence, a truly probabilistic concept. For random variables, the concept reduces to the earlier definition: they are independent if and only if their joint distribution is the product of their marginal distributions.

Throughout, $(\Omega, \mathcal{H}, \mathbb{P})$ is a probability space. As usual, if $\mathcal{G}$ is a sub-σ-algebra of $\mathcal{H}$, we regard it both as a collection of events and as the collection of all numerical random variables that are measurable with respect to it. Recall that σX is the σ-algebra on Ω generated by X, and X here can be a random variable or a collection of random variables. Finally, we write $\mathcal{F}_I$ for $\bigvee_{i\in I} \mathcal{F}_i$ as in I.1.8 and refer to it as the σ-algebra generated by the collection of σ-algebras $\mathcal{F}_i$, $i \in I$.

Definitions

For a fixed integer $n \geq 2$, let $\mathcal{F}_1, \ldots \mathcal{F}_n$ be sub-σ-algebras of $\mathcal{H}$. Then, $\{\mathcal{F}_1, \ldots \mathcal{F}_n\}$ is called an *independency* if

$$\mathbb{E}\ V_1 \cdots V_n = \mathbb{E}V_1 \cdots \mathbb{E}V_n \tag{5.1}$$

for all positive random variables $V_1, \ldots, V_n$ in $\mathcal{F}_1, \ldots, \mathcal{F}_n$ respectively. The term "independency" is meant to suggest a realm governed by the independence of its constituents.

Let T be an arbitrary index set. Let $\mathcal{F}_t$ be a sub-σ-algebra of $\mathcal{H}$ for each t in T. The collection $\{\mathcal{F}_t : \ t \in T\}$ is called an *independency* if its every finite subset is an independency.

In general, elements of an independency are said to be *independent*, or *mutually independent* if emphasis is needed. In loose language, given some objects, the objects are said to be *independent* if the σ-algebras generated by those objects are independent. The objects themselves can be events, random variables, collections of random variables, σ-algebras on Ω, or collections of such, and so on, and they might be mixed. For example, a random variable

X and a stochastic process $\{Y_t : t \in T\}$ and a collection $\{\mathcal{F}_i : i \in I\}$ of σ-algebras on Ω are said to be *independent* if

$$\mathcal{G}_1 = \sigma X\,, \quad \mathcal{G}_2 = \sigma\{Y_t : t \in T\}\,, \quad \mathcal{G}_3 = \mathcal{F}_I = \bigvee_{i\in I} \mathcal{F}_i$$

are independent, that is, if $\{\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3\}$ is an independency.

Independence of σ-algebras

Since a collection of sub-σ-algebras of $\mathcal{H}$ is an independency if and only if its every finite subset is an independency, we concentrate on the independence of a finite number of sub-σ-algebras of $\mathcal{H}$. We start with a test for independence.

5.2 Proposition. *Let $\mathcal{F}_1, \dots, \mathcal{F}_n$ be sub-σ-algebras of $\mathcal{H}$, $n \geq 2$. For each $i \leq n$, let $\mathcal{C}_i$ be a p-system that generates $\mathcal{F}_i$. Then, $\mathcal{F}_1, \dots, \mathcal{F}_n$ are independent if and only if*

$$\mathbb{P}(H_1 \cap \cdots \cap H_n) = \mathbb{P}(H_1) \cdots \mathbb{P}(H_n) \tag{5.3}$$

for all H_i in $\bar{\mathcal{C}}_i = \mathcal{C}_i \cup \{\Omega\}$, $i = 1, \dots, n$.

Proof. Necessity is obvious: take the V_i in 5.1 to be the indicators of the events H_i. To show the sufficiency part, assume 5.3 for H_i in $\bar{\mathcal{C}}_i$, $i = 1, \dots, n$. Fix $H_2, \dots, H_n$ in $\bar{\mathcal{C}}_2, \dots, \bar{\mathcal{C}}_n$ respectively, and let $\mathcal{D}$ be the set of all events H_1 in $\mathcal{F}_1$ for which 5.3 holds. By assumption, $\mathcal{D} \supset \mathcal{C}_1$ and $\Omega \in \mathcal{D}$, and the other two conditions for $\mathcal{D}$ to be a d-system on Ω are checked easily. It follows from the monotone class theorem that $\mathcal{D} \supset \sigma\mathcal{C}_1 = \mathcal{F}_1$. Repeating the procedure successively with $H_2, \dots, H_n$ we see that 5.3 holds for all $H_1, \dots, H_n$ in $\mathcal{F}_1, \dots, \mathcal{F}_n$ respectively. In other words, 5.1 holds when the V_i are indicators. This is extended to arbitrary positive random variables V_i in $\mathcal{F}_i$ by using the form $V_i = \sum_{j=1}^{\infty} a_{ij} 1_{H_{ij}}$ (see Exercise I.2.27) and applying the monotone convergence theorem repeatedly. □

Independence of collections

The next proposition shows that independence survives groupings.

5.4 Proposition. *Every partition of an independency is an independency.*

Proof. Let $\{\mathcal{F}_t : t \in T\}$ be an independency. Let $\{T_1, T_2, \dots\}$ be a partition of T. Then, the subcollections $\mathcal{F}_{T_i} = \{\mathcal{F}_t : t \in T_i\}$, $i \in \mathbb{N}^*$, form a partition of the original independency. The claim is that they are independent, that is, $\{\mathcal{F}_{T_1}, \dots, \mathcal{F}_{T_n}\}$ is an independency for each n. This follows from the preceding proposition: let $\mathcal{C}_i$ be a p-system of all events having the form of an intersection of finitely many events chosen from $\bigcup_{t\in T_i} \mathcal{F}_t$. Then, $\mathcal{C}_i$ generates $\mathcal{F}_{T_i}$ and $\Omega \in \mathcal{C}_i$, and 5.3 holds for the elements of $\mathcal{C}_1, \dots, \mathcal{C}_n$ by the independence of the $\mathcal{F}_t$, $t \in T$. Thus, $\mathcal{F}_{T_1}, \dots, \mathcal{F}_{T_n}$ are independent, and this is for arbitrary n. □

Pairwise independence

A collection of objects (like σ-algebras, random variables) are said to be *pairwise independent* if every pair of them is an independency. This is, of course, much weaker than being mutually independent. But independence can be checked by repeated checks for pairwise independence. We state this for a sequence of σ-algebras; it holds for a finite sequence as well, and therefore can be used to check the independency for arbitrary collections.

5.5 PROPOSITION. *The sub-σ-algebras $\mathcal{F}_1, \mathcal{F}_2, \ldots$ of $\mathcal{H}$ are independent if and only if $\mathcal{F}_{\{1,\ldots,n\}}$ and $\mathcal{F}_{n+1}$ are independent for all $n \geq 1$.*

Proof. Necessity is immediate from the last proposition. For sufficiency, suppose that $\mathcal{G}_n = \mathcal{F}_{\{1,\ldots,n\}} = \bigvee_{i=1}^{n} \mathcal{F}_i$ and $\mathcal{F}_{n+1}$ are independent for all n. Then, for $H_1, \ldots, H_m$ in $\mathcal{F}_1, \ldots, \mathcal{F}_m$ respectively, we can see that 5.3 holds by repeated applications of the independence of $\mathcal{G}_n$ and $\mathcal{F}_{n+1}$ for $n = m-1$, $m-2, \ldots, 1$ in that order. Thus, $\mathcal{F}_1, \ldots, \mathcal{F}_m$ are independent by Proposition 5.2, and this is true for all $m \geq 2$. □

Independence of random variables

For each t in some index set T, let X_t be a random variable taking values in some measurable space $(E_t, \mathcal{E}_t)$. According to the general definitions above, the variables X_t are said to be *independent*, and the collection $\{X_t : t \in T\}$ is called an *independency*, if $\{\sigma X_t : t \in T\}$ is an independency.

Since a collection is an independency if and only if its every finite subset is an independency, we concentrate on the independence of a finite number of them, which amounts to taking $T = \{1, 2, \ldots, n\}$ for some integer $n \geq 2$.

5.6 PROPOSITION. *The random variables $X_1, \ldots, X_n$ are independent if and only if*

$$\mathbb{E}\ f_1 \circ X_1 \cdots f_n \circ X_n = \mathbb{E}\ f_1 \circ X_1 \cdots \mathbb{E}\ f_n \circ X_n \tag{5.7}$$

for all positive functions $f_1, \ldots, f_n$ in $\mathcal{E}_1, \ldots, \mathcal{E}_n$ respectively.

Proof. We need to show that 5.1 holds for all positive $V_1, \ldots, V_n$ in $\sigma X_1, \ldots, \sigma X_n$ respectively if and only if 5.7 holds for all positive $f_1, \ldots, f_n$ in $\mathcal{E}_1, \ldots, \mathcal{E}_n$ respectively. But this is immediate from Theorem 4.4: $V_i \in \sigma X_i$ if and only if $V_i = f_i \circ X_i$ for some f_i in $\mathcal{E}_i$. □

Let π be the joint distribution of $X_1, \ldots, X_n$, and let $\mu_1, \ldots, \mu_n$ be the corresponding marginals. Then, the left and the right sides of 5.7 are equal to, respectively,

$$\int_{E_1 \times \cdots \times E_n} \pi(dx_1, \ldots, dx_n)\, f_1(x_1) \cdots f_n(x_n)$$

and

$$\int_{E_1} \mu_1(dx_1)\, f_1(x_1) \int_{E_2} \cdots \int_{E_n} \mu_n(dx_n)\, f_n(x_n)$$

The equality of these two expressions for all positive $f_1, \dots, f_n$ is equivalent to saying that $\pi = \mu_1 \times \cdots \times \mu_n$. We state this next.

5.8 PROPOSITION. *The random variables $X_1, \dots, X_n$ are independent if and only if their joint distribution is the product of their marginal distributions.*

Finally, a comment on functions of independent variables. In the language of Exercise 4.14, let $Y_1, \dots, Y_n$ be determined by $X_1, \dots, X_n$ respectively. Then $\sigma Y_i \subset \sigma X_i$ for $i = 1, \dots, n$, and it follows from the definition of independency that $Y_1, \dots, Y_n$ are independent if $X_1, \dots, X_n$ are independent. We state this observation next.

5.9 PROPOSITION. *Measurable functions of independent random variables are independent.*

Sums of independent random variables

Let X and Y be $\mathbb{R}^d$-valued independent random variables with distributions μ and ν respectively. Then, the distribution of (X, Y) is the product measure $\mu \times \nu$, and the distribution $\mu * \nu$ of $X + Y$ is given by

$$5.10 \qquad (\mu * \nu) f = \mathbb{E} f(X + Y) = \int_{\mathbb{R}} \mu(dx) \int_{\mathbb{R}} \nu(dy)\, f(x + y),$$

This distribution $\mu * \nu$ is called the *convolution* of μ and ν. See exercises below for more. Of course, since $X + Y = Y + X$, we have $\mu * \nu = \nu * \mu$. The convolution operation can be extended to any number of distributions.

Sums of random variables and the limiting behavior of such sums as the number of summands grows to infinity are of constant interest in probability theory. We shall return to such matters repeatedly in the chapters to follow. For the present, we describe two basic results, zero-one laws due to Kolmogorov and Hewitt-Savage.

Kolmogorov's 0-1 law

Let $(\mathcal{G}_n)$ be a sequence of sub-σ-fields of $\mathcal{H}$. We think of $\mathcal{G}_n$ as the information revealed by the n^{th} trial of an experiment. Then, $\mathcal{T}_n = \bigvee_{m>n} \mathcal{G}_m$ is the information about the future after n, and $\mathcal{T} = \bigcap_n \mathcal{T}_n$ is that about the remote future. The last is called the *tail-σ-algebra*; it consists of events whose occurrences are unaffected by the happenings in finite time.

5.11 EXAMPLE. Let $X_1, X_2, \dots$ be real valued random variables, put $\mathcal{G}_n = \sigma X_n$ and $S_n = X_1 + \cdots + X_n$.

a) The event $\{\omega : \lim_n S_n(\omega) \text{ exists}\}$ belongs to $\mathcal{T}_n$ for every n and, hence, belongs to the tail-σ-algebra $\mathcal{T}$.

b) Similarly, $\{\limsup \frac{1}{n} S_n > b\}$ is unaffected by the first n variables, and this is true for all n, and hence this event belongs to $\mathcal{T}$.

c) But, $\{\limsup S_n > b\}$ is not in $\mathcal{T}$.

d) Let B be a Borel subset of $\mathbb{R}$. Let $\{X_n \in B \text{ i.o.}\}$, read X_n is in B *infinitely often*, be the set of ω for which $\sum_n 1_B \circ X_n(\omega) = +\infty$. This event belongs to $\mathcal{T}$.

e) The event $\{S_n \in B \text{ i.o.}\}$ is not in $\mathcal{T}$.

The following theorem, called Kolmogorov's 0-1 law, implies in particular that, if the X_n of the preceding example are independent, then each one of the events in $\mathcal{T}$ has probability equal to either 0 or 1.

5.12 THEOREM. *Let $\mathcal{G}_1, \mathcal{G}_2, \ldots$ be independent. Then, $\mathbb{P}(H)$ is either 0 or 1 for every event H in the tail $\mathcal{T}$.*

Proof. By Proposition 5.4 on partitions of independencies, $\{\mathcal{G}_1, \ldots, \mathcal{G}_n, \mathcal{T}_n\}$ is an independency for every n, which implies that so is $\{\mathcal{G}_1, \ldots, \mathcal{G}_n, \mathcal{T}\}$ for every n, since $\mathcal{T} \subset \mathcal{T}_n$. Thus, by definition, $\{\mathcal{T}, \mathcal{G}_1, \mathcal{G}_2, \ldots\}$ is an independency, and so is $\{\mathcal{T}, \mathcal{T}_0\}$ by Proposition 5.4 again. In other words, for H in $\mathcal{T}$ and $G \in \mathcal{T}_0$, we have $\mathbb{P}(H \cap G) = \mathbb{P}(H) \cdot \mathbb{P}(G)$, and this holds for $G = H$ as well because $\mathcal{T} \subset \mathcal{T}_0$. Thus, for H in $\mathcal{T}$, we have $\mathbb{P}(H) = \mathbb{P}(H) \cdot \mathbb{P}(H)$, which means that $\mathbb{P}(H)$ is either 0 or 1. □

As a corollary, assuming that the $\mathcal{G}_n$ are independent, for every random variable V in the tail-σ-algebra there is a constant c in $\bar{\mathbb{R}}$ such that $V = c$ almost surely. Going back to Example 5.11, for instance, $\limsup S_n/n$ is almost surely constant. In the same example, the next theorem will imply that the events $\{\limsup S_n > b\}$ and $\{S_n \in B \text{ i.o.}\}$ have probability 0 or 1, even though they are not in the tail $\mathcal{T}$, provided that we add to the independence of X_n the extra condition that they have the same distribution.

Hewitt-Savage 0-1 law

Let $X = (X_1, X_2, \ldots)$, where the X_n take values in some measurable space $(E, \mathcal{E})$. Let $\mathcal{F} = (\mathcal{F}_1, \mathcal{F}_2, \ldots)$ be the filtration generated by X, that is, $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$ for each n. Put $\mathcal{F}_\infty = \lim \mathcal{F}_n = \bigvee_n \mathcal{F}_n$, and recall from Theorem 4.4 and its sequel that $\mathcal{F}_\infty$ consists of random variables of the form $V = f \circ X$ with f in $\mathcal{E}^\infty$, and $\mathcal{F}_n$ consists of the variables of the form $V_n = f_n(X_1, \ldots, X_n) = \hat{f}_n \circ X$ with f_n in $\mathcal{E}^n$ (and appropriately defined $\hat{f}_n$).

By a *finite permutation* p is meant a bijection $p : \mathbb{N}^* \mapsto \mathbb{N}^*$ such that $p(n) = n$ for all but finitely many n. For such a permutation p, we write

$$X \circ p = (X_{p(1)}, X_{p(2)}, \ldots), \tag{5.13}$$

which is a re-arrangement of the entries of X. The notation extends to arbitrary random variables V in $\mathcal{F}_\infty$: if $V = f \circ X$ then $V \circ p = f \circ (X \circ p)$. It will

be useful to note that, if the X_n are independent and identically distributed, the probability laws of X and $X \circ p$ are the same, and hence, the distributions of V and $V \circ p$ are the same.

A random variable V in $\mathcal{F}_\infty$ is said to be *permutation invariant* if $V \circ p = V$ for every finite permutation p. An event in $\mathcal{F}_\infty$ is said to be permutation invariant if its indicator is such. These are variables like $V = \limsup S_n$ or events like $\{S_n \in B \text{ i.o.}\}$ in Example 5.11; they are unaffected by the re-arrangements of the entries of X by finite permutations.

The collection of all permutation invariant events is a σ-algebra which contains the tail-σ-algebra of X. The following, called Hewitt-Savage 0-1 law, shows that it is almost surely trivial (just as the tail) provided that the X_n are identically distributed in addition to being independent.

5.14 THEOREM. *Suppose that $X_1, X_2, \ldots$ are independent and identically distributed. Then, every permutation invariant event has probability 0 or 1. Also, for every permutation invariant random variable V there is a constant c in $\bar{\mathbb{R}}$ such that $V = c$ almost surely.*

Proof. It is sufficient to show that if $V : \Omega \mapsto [0,1]$ is a permutation invariant variable in $\mathcal{F}_\infty$, then $\mathbb{E}(V^2) = (\mathbb{E}V)^2$. Let V be such. By Proposition 4.9 there are V_n in $\mathcal{F}_n$, $n \geq 1$, such that each V_n takes values in $[0,1]$ and

$$\lim \mathbb{E}|V - V_n| = 0 \qquad \lim_n \mathbb{E}V_n = \mathbb{E}V, \tag{5.15}$$

the second limit being a consequence of the first.

Fix n. Let p be a finite permutation. The assumption about X implies that X and $X \circ p$ have the same probability law, which in turn implies that U and $U \circ p$ have the same distribution for every U in $\mathcal{F}_\infty$. Taking $U = V - V_n$, noting that $U \circ p = V \circ p - V_n \circ p = V - V_n \circ p$ by the invariance of V, we see that

$$\mathbb{E}|V - V_n \circ p| = \mathbb{E}|V - V_n|. \tag{5.16}$$

This is true, in particular, for the permutation $\hat{p}$ that maps $1, \ldots, n$ to $n+1, \ldots, 2n$ and vice-versa, leaving $\hat{p}(m) = m$ for $m > 2n$. We define $\hat{V}_n = V_n \circ \hat{p}$ and observe that, if $V_n = f_n(X_1, \ldots, X_n)$, then $\hat{V}_n = f_n(X_{n+1}, \ldots, X_{2n})$, which implies that V_n and $\hat{V}_n$ are independent and have the same distribution. Together with 5.16, this yields

$$\mathbb{E}V_n\hat{V}_n = (\mathbb{E}V_n)^2\,, \qquad \mathbb{E}|V - \hat{V}_n| = \mathbb{E}|V - V_n|, \tag{5.17}$$

which in turn show that

$$|\mathbb{E}(V^2) - (\mathbb{E}V_n)^2| = |\mathbb{E}(V^2 - V_n\hat{V}_n)| \leq \mathbb{E}|V^2 - V_n\hat{V}_n| \leq 2\mathbb{E}|V - V_n|, \tag{5.18}$$

where the final step used (recalling $|V| \leq 1$ and $|V_n| \leq 1$)

$$|V^2 - V_n\hat{V}_n| = |(V - V_n)V + (V - \hat{V}_n)V_n| \leq |V - V_n| + |V - \hat{V}_n|,$$

and 5.17. Applying 5.15 to 5.18 yields the desired result that $\mathbb{E}V^2 = (\mathbb{E}V)^2$. □

5.19 Example. *Random walks.* This is to provide a typical application of the preceding theorem. Returning to Example 5.11, assume further that $X_1, X_2, \ldots$ have the same distribution. Then, the stochastic process (S_n) is called a *random walk* on $\mathbb{R}$. The avoid the trivial case where $S_1 = S_2 = \ldots = 0$ almost surely, we assume that $\mathbb{P}\{X_1 = 0\} < 1$. Then, concerning the limiting behavior of the random walk, there are three possibilities, exactly one of which is almost sure:

i) $\lim S_n = +\infty$,

ii) $\lim S_n = -\infty$,

iii) $\liminf S_n = -\infty$, and $\limsup S_n = +\infty$.

Here is the argument for this. By the preceding theorem, there is a constant c in $\bar{\mathbb{R}}$ such that $\limsup S_n = c$ almost surely. Letting $\hat{S}_n = S_{n+1} - X_1$ yields another random walk $(\hat{S}_n)$ which has the same law as (S_n). Thus, $\limsup \hat{S}_n = c$ almost surely, which means that $c = c - X_1$. Since we excluded the trivial case when $\mathbb{P}\{X_1 = 0\} = 1$, it follows that c is either $+\infty$ or $-\infty$. Similarly, $\liminf S_n$ is either almost surely $-\infty$ or almost surely $+\infty$. Of the four combinations, discarding the impossible case when $\liminf S_n = +\infty$ and $\limsup S_n = -\infty$, we arrive at the result.

If the common distribution of the X_n is *symmetric*, that is, if X_1 and $-X_1$ have the same distribution (like the Gaussian with mean 0), then (S_n) and $(-S_n)$ have the same law, and it follows that the cases (i) and (ii) are improbable. So then, case (iii) holds almost surely.

Exercises

5.20 *Independence and functional independence.* Suppose that $(\Omega, \mathcal{H}, \mathbb{P}) = (\mathbb{B}, \mathcal{B}, \lambda) \times (\mathbb{B}, \mathcal{B}, \lambda)$, where $\mathbb{B} = [0,1]$, $\mathcal{B} = \mathcal{B}(\mathbb{B})$ and λ is the Lebesgue measure on $\mathbb{B}$. For each $\omega = (\omega_1, \omega_2)$ in Ω, let $X(\omega) = f(\omega_1)$ and $Y(\omega) = g(\omega_2)$ for some Borel functions f and g on $\mathbb{B}$. Show that X and Y are independent.

5.21 *Independence and transforms.* Let X and Y be positive random variables. Then, X and Y are independent if and only if their joint Laplace transform is the product of their Laplace transforms, that is, if and only if

$$\mathbb{E}e^{-pX-qY} = \mathbb{E}e^{-pX}\,\mathbb{E}e^{-qY}\,, \qquad p, q \in \mathbb{R}_+.$$

Show this recalling that the joint Laplace transforms determine the joint distributions. A similar result holds for X and Y real-valued, but with characteristic functions. Obviously, these results can be extended to any finite number of variables.

5.22 *Sums of independent variables.* Let X and Y be independent real-valued random variables. Show that the characteristic function of $X + Y$ is the product of their characteristic functions. When X and Y are positive, the

same is true with Laplace transforms. When X and Y are positive integers, the same holds with generating functions. Use these to show the following.

a) If X has the Poisson distribution with mean a, and Y the Poisson distribution with mean b, then $X+Y$ has the Poisson distribution with mean $a+b$.

b) If X has the gamma distribution with shape index a and scale parameter c, and Y has the gamma distribution with shape index b and the same scale parameter c, then $X+Y$ has the gamma distribution with shape index $a+b$ and scale c.

c) If X has the Gaussian distribution with mean a and variance b and Y has the Gaussian distribution with mean c and variance d, then $X+Y$ has the Gaussian distribution with mean $a+c$ and variance $b+d$.

5.23 *Convolutions*

a) Let μ and ν be probability measures on $\mathbb{R}$, and let $\pi = \mu * \nu$ be defined by 5.10. Show that

$$\pi(B) = \int_{\mathbb{R}} \mu(dx)\, \nu(B-x)\ , \qquad B \in \mathcal{B}_{\mathbb{R}},$$

where $B - x = \{y - x :\ y \in B\}$.

b) Let λ be the Lebesgue measure on $\mathbb{R}$. Suppose that $\mu(dx) = \lambda(dx)\, p(x)$ and $\nu(dx) = \lambda(dx)\, q(x)$, $x \in \mathbb{R}$, for some positive Borel functions p and q. Show that, then, $\pi(dx) = \lambda(dx)\, r(x)$, where

$$r(x) = \int_{\mathbb{R}} dy\, p(y)\, q(x-y)\ , \qquad x \in \mathbb{R}.$$

Historically, then, r is said to be the convolution of the functions p and q, and the notation $r = p * q$ is used to indicate it.

c) Let μ and ν be as in the preceding case, but be carried by $\mathbb{R}_+$. Then, p and q vanish outside $\mathbb{R}_+$, and

$$r(x) = \int_0^x dy\, p(y)\, q(x-y)\ , \qquad x \in \mathbb{R}_+,$$

with $r(x) = 0$ for x outside $\mathbb{R}_+$.

Complements: Bernoulli sequences

5.24 *Bernoulli variables.* These are random variables that take the values 0 and 1 only. Each such variable is the indicator of an event, the event being named "success" to add distinction. Thus, if X is a Bernoulli variable, $p = \mathbb{P}\{X = 1\}$ is called the success probability, and then, $q = \mathbb{P}\{X = 0\} = 1 - p$ becomes the failure probability. Show that

$$\mathbb{E}\ X = \mathbb{E}\ X^2 = \cdots = p, \quad \text{Var}\ X = pq, \quad \mathbb{E}\ z^X = q + pz.$$

5.25 *Bernoulli trials.* Let $X_1, X_2, \ldots$ be Bernoulli variables. It is usual to think of X_n as indicating the result of the n^{th} trial in a sequence of trials: $X_n(\omega) = 1$ means that a "success" has occurred at the n^{th} trial corresponding to the sequence described by the outcome ω. Often, it is convenient to assume that the trials occur at times $1, 2, 3, \ldots$. Then,

$$S_n = X_1 + \cdots + X_n$$

is the number of successes occuring during the time interval $[1, n]$. Assuming that $X_1, X_2, \ldots$ are independent and have the same success probability p (and the same failure probability $q = 1 - p$), show that

$$\mathbb{P}\{S_n = k\} = \frac{n!}{k!(n-k)!} p^k q^{n-k}, \quad k = 0, 1, \ldots, n.$$

Hint: First compute $\mathbb{E}\, z^{S_n}$ using 5.24, and recall the binomial expansion $(a+b)^n = \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} a^k b^{n-k}$. For this reason, the distribution of S_n is called the *binomial distribution.*

5.26 *Times of successes.* Let $X_1, X_2, \ldots$ be independent Bernoulli variables with the same success probability p. Define, for each k in $\mathbb{N}^*$, the time of k^{th} success by

$$T_k(\omega) = \inf\{n \geq 1 : \ S_n(\omega) \geq k\}, \quad \omega \in \Omega.$$

Note that this yields $T_k(\omega) = +\infty$ if $S_n(\omega) < k$ for all n. Show that T_k is a random variable for each k in $\mathbb{N}^*$. Show that, for integers $n \geq k$,

$$\mathbb{P}\{T_k = n\} = \frac{(n-1)!}{(k-1)!(n-k)!} p^k q^{n-k}, \quad \mathbb{P}\{T_k \leq n\} = \sum_{j=k}^{n} \frac{n!}{j!(n-j)!} p^j q^{n-j}.$$

Show, in particular, that $T_k < \infty$ almost surely and, therefore, that $\lim S_n = +\infty$ almost surely.

5.27 *Waits between successes.* Let the X_n be as in 5.26. For $k \in \mathbb{N}^*$, define the waiting time $W_k(\omega)$ between the $(k-1)^{\text{th}}$ and k^{th} successes by letting $T_0(\omega) = 0$ and

$$W_k(\omega) = \begin{cases} T_k(\omega) - T_{k-1}(\omega) & \text{if } T_k(\omega) < \infty, \\ +\infty & \text{otherwise.} \end{cases}$$

For integers $i_1, \ldots, i_k$ in $\mathbb{N}^*$, express the event $\{W_1 = i_1, \ldots, W_k = i_k\}$ in terms of the variables X_k, compute the probability of the event in question, and conclude that $W_1, W_2, \ldots, W_k$ are independent random variables with the same distribution

$$\mathbb{P}\{W_k = i\} = pq^{i-1}, \quad i \in \mathbb{N}^*.$$

This distribution on $\mathbb{N}^*$ is called the *geometric distribution* with success probability p. Compute

$$\mathbb{E}\, W_k, \ \text{Var}\, W_k, \ \mathbb{E}\, T_k, \ \text{Var}\, T_k.$$

5.28 *Multinomial trials.* Let $X_1, X_2, \ldots$ be mutually independent random variables taking values in a finite set D, say $D = \{a, \ldots, d\}$, with

$$\mathbb{P}\{X_n = x\} = p(x), \quad x \in D.$$

For each x in D and ω in Ω, let $S_n(\omega, x)$ be the number of times that x appears in $(X_1(\omega), \ldots, X_n(\omega))$. Then, $S_n(x) : \omega \mapsto S_n(\omega, x)$ is a random variable for each n in $\mathbb{N}^*$ and each point x in D. Show that

$$\mathbb{P}\{S_n(a) = k(a), \ldots, S_n(d) = k(d)\} = \frac{n!}{k(a)! \cdots k(d)!} p(a)^{k(a)} \cdots p(d)^{k(d)}$$

for all $k(a), \ldots, k(d)$ in $\mathbb{N}$ with $k(a) + \cdots + k(d) = n$. This defines a probability measure on the simplex of all vectors $(k(a), \ldots, k(d))$ with $k(a) + \cdots + k(d) = n$; it is called a *multinomial distribution.*

5.29 *Empirical distributions.* Let $X_1, X_2, \ldots$ be mutually independent random variables taking values in some measurable space $(E, \mathcal{E})$and having the same distribution μ. Define

$$S_n(\omega, A) = \sum_{i=1}^{n} 1_A \circ X_i(\omega), \quad n \in \mathbb{N}, \omega \in \Omega, A \in \mathcal{E}.$$

Then, $A \mapsto S_n(\omega, A)$ is a counting measure on $(E, \mathcal{E})$whose atoms are the locations $X_1(\omega), \ldots, X_n(\omega)$, and $\frac{1}{n} S_n(\omega, A)$ defines a probability measure on $(E, \mathcal{E})$, called the *empirical distribution* corresponding to $X_1(\omega), \ldots, X_n(\omega)$. Writing $S_n(A)$ for the random variable $\omega \mapsto S_n(\omega, A)$, show that

$$\mathbb{P}\{S_n(A_1) = k_1, \ldots, S_n(A_m) = k_m\} = \frac{n!}{k_1! \cdots k_m!} \mu(A_1)^{k_1} \cdots \mu(A_m)^{k_m}$$

for every measurable partition $(A_1, \ldots, A_m)$ of E and integers $k_1, \ldots, k_m \geq 0$ summing to n.

5.30 *Inclusion-exclusion principle.* Let $X_1, \ldots, X_j$ be Bernoulli variables. Show that

$$\mathbb{P}\{X_1 = \cdots = X_j = 0\} = \mathbb{E} \sum_{Y} Y_1 \cdots Y_j$$

where the sum is over all j-tuples $Y = (Y_1, \ldots, Y_j)$ with each Y_i being either 1 or $-X_i$. Hint: The left side is the expectation of $(1 - X_1) \cdots (1 - X_j)$.

5.31 *Continuation.* For $X_1, \ldots, X_k$ Bernoulli, show that

$$\mathbb{P}\{X_1 = \cdots = X_j = 0, X_{j+1} = \cdots = X_k = 1\} = \mathbb{E} \sum_{Y} Y_1 \cdots Y_j X_{j+1} \cdots X_k$$

where the sum is over all Y as in 5.30.

5.32 *Probability law of a collection of Bernoullis.* Let I be an arbitrary index set. For each i in I, let X_i be a Bernoulli variable. Show that the probability law of $X = \{X_i : i \in I\}$ is specified by

$$\mathbb{E} \prod_{i \in J} X_i, \quad J \subset I, \ J \text{ finite} ;$$

in other words, knowing these expectations is enough to compute

$$\mathbb{P}\{X_i = b_i, i \in K\}$$

for every finite subset K of I and binary numbers b_i, $i \in K$.

5.33 *Independence of Bernoullis.* Show that X in 5.32 is an independency if and only if, for every finite $J \subset I$,

$$\mathbb{E} \prod_{i \in J} X_i = \prod_{i \in J} \mathbb{E}\ X_i.$$