SMSTC (2007/08)

Probability

www.smstc.ac.uk

Contents

SMSTC (2007/08)

Probability

3: Random variables

Takis Konstantopoulos[a]

www.smstc.ac.uk

Contents

3.1 Introduction

Random variables are at the core of Probability and Modern Mathematics and have nothing to do with randomness: they are merely functions, labels of outcomes of an "abstract space" Ω. These functions must be compatible with the set of events of Ω. This compatibility property is called measurability. The concept of a random variable is purely functional.

David Mumford [4] challenges the mathematical community by insisting that we are encountering a paradigm shift in Mathematics, namely, we are seeing how "Stochasticity" plays a much more fundamental role within Mathematics than previously envisioned. Moreover, Mumford claims that random variables should be put first in the discipline of Probability. Somehow, he suggests a change of the neo-classical Kolmogorov foundation, and wants to replace it by functions, i.e. random variables, as primary objects. There is no doubt that this can be done, just as Topology can be approached, not with sets, but with functions to start with. While we will not be as "radical" as Mumford, we will explain the importance of random variables by explicitly discussing all (classical) nuances associated with them.

So we first start with functional concepts, ignoring probability. We then put a probability and see how it is transformed by a random variable: a random variable transforms a probability into another probability that is known as its law.

Discrete random variables are trivial, from a foundational point of view (but far from trivial otherwise). Continuous random variables are hard to visualise. We prefer to construct such random variables. by flipping coins which is the most natural thing to do in Probability (arguably the only thing to do).

The distribution of a random variable is any function that uniquely specifies its law. It is a convention, for real random variables, to use the so-called distribution function as a distribution. Some distribution functions are nice in the sense that they can be differentiated and their derivatives are useful because they can be integrated to enable us to compute probabilities. These derivatives are called densities. We try to be semi-honest when discussing them.

The expectation of a random variable is discussed, constructed, and shown to be compatible with the usual naïve formulae.

Some *sine qua non* inequalities (under the names of Markov, Chebyshev, Chernoff, Jensen, Cauchy, Bunyakowskii, Schwarz, Hölder and Minkowski) are presented, because inequalities are more useful than equalities.

Finally, there is a brief discussion of he concept of a moment generating function.

Exercises are scattered within the text.

Appendix A summarises basic notions regarding sets. Appendix B is gives a somewhat elementary approach to putting probabilities on sets.

3.2 Random variables are functions

A measurable space is a set together with a σ-field of subsets of it. The concept of a random variable was introduced in $\boxed{1.1.5}$. Intuitively, a random variable assigns concrete labels to abstract outcomes. Concretely,

> A RANDOM VARIABLE IS A MEASURABLE FUNCTION X between two measurable spaces, an "abstract" one, $(\Omega, \mathscr{F})$, and a "concrete"[a] one, $(S, \mathscr{S})$.

[a] The adjectives in quotes have nothing to do with Mathematics but, rather, with our human interpretation of it.

In other words, we require that the inverse image by X of each element of $\mathscr{S}$ be an element of $\mathscr{F}$. We denote this situation by

$$X : (\Omega, \mathscr{F}) \to (S, \mathscr{S}).$$

Many a times $(\Omega, \mathscr{F})$ could (or should) be left unspecified but $(S, \mathscr{S})$ should be chosen specifically; for example it could be $(\mathbb{R}, \mathscr{B})$, where $\mathscr{B} = \mathscr{B}(\mathbb{R})$ is the class of Borel sets (see $\boxed{1.1.3}$) on $\mathbb{R}$. In such a case we call X ONE (REAL) RANDOM VARIABLE, omitting the numeral ONE when not necessary. As another example, let $(S, \mathscr{S}) = (\mathbb{R}^2, \mathscr{B}(\mathbb{R}^2))$, where $\mathscr{B}(\mathbb{R}^2)$ are the Borel sets on $\mathbb{R}^2$ (defined as the smallest σ-field containing all open sets in $\mathbb{R}^2$). In this case we refer to X as TWO (REAL) RANDOM VARIABLES, because we may, by choosing Cartesian coördinates on $\mathbb{R}^2$, represent X by (X_1, X_2), where X_1 is one random variable and X_2 is also one random variable. More generally, we may let $(S, \mathscr{S}) = (\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$, and then we refer to X a RANDOM (d-DIMENSIONAL REAL) VECTOR.

Terminology: we say that X is a random variables IN S or that X is a random element OF S. If S is a concrete space with a name "abc" then we call X a random "abc". For instance, if S is the space of triangles on the plane, then a random variable in X is called random triangle.

EXERCISE 1. Show that if $X : (\Omega, \mathscr{F}) \to (S, \mathscr{S})$, $H : (S, \mathscr{S}) \to (T, \mathscr{T})$ are random variables then $H \circ X$ is a random variable.

The following lemma, useful in checking that a function is a random variable:

Lemma 3.1. *If $X : \Omega \to S$ is a function, $\mathscr{F}$ a σ-field on Ω and $\mathscr{S}$ a σ-field on S generated by the collection of sets $\mathscr{C}$ then X is a random variable if and only if $X^{-1}(B) \in \mathscr{F}$ for all $B \in \mathscr{C}$.*

Proof Let $\mathscr{S}' := \{B \subset S : X^{-1}(B) \in \mathscr{F}\}$. This is a σ-field. Indeed, if $B \in \mathscr{S}'$ then $X^{-1}(B^c) = X^{-1}(B)^c \in \mathscr{F}$ because $X^{-1}(B) \in \mathscr{F}$. If $B_1, B_2, \ldots \in \mathscr{S}'$ then $X^{-1}(\cap_j B_j) = \cap_j X^{-1}(B_j) \in \mathscr{F}$ because each $X^{-1}(B_j) \in \mathscr{F}$. Suppose that $X^{-1}(B) \in \mathscr{F}$ for all $B \in \mathscr{C}$. This means that $\mathscr{C}$ is contained in $\mathscr{S}'$. But then $\mathscr{S}$ which is the smallest σ-field containing $\mathscr{C}$ must be contained in $\mathscr{S}'$ for the latter is a σ-field. In symbols, $\mathscr{S} \subset \mathscr{S}'$. But, by definition of $\mathscr{S}'$, this means that $X^{-1}(B) \in \mathscr{F}$ for any $B \in \mathscr{S}$. □

Recall that the notation

$$\{X \in B\} = X^{-1}(B)$$

is used all the time, so we will stick to it when we can.

Corollary 3.1. *If $X : \Omega \to \mathbb{R}$ is a function and $\mathscr{F}$ a σ-field on Ω, then X is one random variable $\iff \{X \le x\} \in \mathscr{F}$ for all $x \in \mathbb{R} \iff \{X < x\} \in \mathscr{F}$ for all $x \in \mathbb{R} \iff \{X > x\} \in \mathscr{F}$ for all $x \in \mathbb{R}$.*

The real fun starts when $(S, \mathscr{S})$ is chosen to be a large (but concrete) space, for example, S can be a subset of a space of real-valued functions on the real line and $\mathscr{S}$ a suitable σ-field. In such a case, X is called RANDOM FUNCTION or STOCHASTIC PROCESS (stochastic being a synonym of random and process being a synonym of function; therefore, random process or stochastic function are also acceptable terminologies for the same thing).

3.3 Continuous functions of random variables

Lemma 3.2. *If X is a random vector in $\mathbb{R}^d$ and $f : \mathbb{R}^d \to \mathbb{R}^k$ is a continuous function then $f \circ X$ is a random vector in $\mathbb{R}^k$.*

Proof Let B be an open subset of $\mathbb{R}^k$. Then $f^{-1}(B)$ is an open subset of $\mathbb{R}^d$ because f is continuous. Hence $f^{-1}(B)$ is a Borel subset of $\mathbb{R}^d$ and so f is measurable. Now use Exercise 1. □

An application of this is that all usual algebraic operations on random variables will result random variables. So

$$X_1 + X_2, \quad X_1 \cdot X_2, \quad X_1 \wedge X_2$$

are all random variables.

A more important result is that we can go beyond algebraic operations and retain measurability. So

Lemma 3.3. *If $X_1, X_2, \dots$ are random variables in $\mathbb{R}$ then*

$$\inf_j X_j, \quad \varliminf_j X_j, \quad \sup_j X_j, \quad \varlimsup_j X_j$$

are random variables in $\mathbb{R} \cup \{+\infty, -\infty\}$.

Proof Indeed,

$$\{\inf_j X_j \le a\} = \cap_j \{X_j \le a\}$$

and, since $\{X_j \le a\} \in \mathscr{F}$ for all j, and $\mathscr{F}$ is a σ-field, then the intersection also belongs to $\mathscr{F}$. For the second variable, we simply observe that $\sup_j X_j = -\inf_j(-X_j)$ is also a random variable and so $\varliminf_j X_j = \sup_j \inf_{k \ge j} X_k$ is also a random variable. □

3.4 Induced σ-fields; measurability

Recall, as in 1.1.3, that

> the σ-field generated by a collection, say $\mathscr{A}$ of subsets of Ω, is defined as the intersection of all σ-fields containing $\mathscr{A}$; it is denoted by $\sigma(\mathscr{A})$.

If $X : \Omega \to S$ is a function, and if $\mathscr{S}$ is a fixed σ-field on S, the σ-field GENERATED or INDUCED by X is defined by

$$\sigma(X) := \{X^{-1}(B), B \in \mathscr{S}\}.$$

It is easy to see that

Lemma 3.4. *(i) $\sigma(X)$ is a σ-field, (ii) $\sigma(X)$ is the intersection of all σ-fields $\mathscr{G} \subset \mathscr{F}$ such that $X : (\Omega, \mathscr{G}) \to (S, \mathscr{S})$ is a random variable, (iii) X is a random variable if and only if $\sigma(X) \subset \mathscr{F}$.*

If $Y : \Omega \mapsto S'$ is another function, and if S' is endowed with another fixed σ-field $\mathscr{S}'$, then the notation $\sigma(X, Y)$ stands for the smallest σ-field containing $\sigma(X) = \{X^{-1}(B), B \in \mathscr{S}\}$ and $\sigma(Y) = \{X^{-1}(B), B \in \mathscr{S}'\}$:

$$\sigma(X, Y) := \sigma(\sigma(X) \cup \sigma(Y)) =: \sigma(X) \vee \sigma(Y).$$

(The last bit is just another notation for the same thing.) Let us study random variables that take values in $\mathbb{R}^d$ or, more generally, in a product $S_1 \times \cdots \times S_d$ of sets. Suppose that on each S_i we have a σ-field $\mathscr{S}_i \subset 2^{S_i}$. We first construct a natural σ-field on $S_1 \times \cdots \times S_d$. For each i consider the projection function

$$\pi_i : S_1 \times \cdots \times S_d \to S_i; \quad \pi_i : (s_1, \cdots, s_d) \mapsto s_i.$$

Define

$$\mathscr{S}_1 \otimes \cdots \otimes \mathscr{S}_d := \sigma(\pi_1, \dots, \pi_d).$$

EXERCISE 2. Consider $(S_i, \mathscr{S}_i)$, $i = 1, \ldots, d$. Let $d = 2$ for simplicity. Show that

$$\begin{aligned}\mathscr{S}_1 \otimes \mathscr{S}_2 &= \sigma(\{B_1 \times S_2 : B_1 \in \mathscr{S}_1\} \cup \{S_1 \times B_2 : B_2 \in \mathscr{S}_2\}).\\ &= \sigma(\{B_1 \times B_2 : B_1 \in \mathscr{S}_1, B_2 \in \mathscr{S}_2\}).\end{aligned}$$

Lemma 3.5. *Let $X_i : (\Omega, \mathscr{F}) \to (S_i, \mathscr{S}_i)$, $i = 1, 2, \ldots, d$, be random variables. Let $S = S_1 \times \cdots \times S_d$, $\mathscr{S} = \mathscr{S}_1 \otimes \cdots \otimes \mathscr{S}_d$. Then $(X_1, \ldots, X_d) : (\Omega, \mathscr{F}) \to (S, \mathscr{S})$ is a random variable.*

Proof We just have to check that $(X_1, \ldots, X_d)^{-1}(B) \in \mathscr{F}$ for B ranging in a suitable π-system. We take this π-system (and it is immediate to see that it is so) to be all sets of the form $B_1 \times \cdots \times B_d$, where $B_j \in \mathscr{S}_j$ for all $1 \le j \le d$. Then $(X_1, \ldots, X_d)^{-1}(B_1 \times \cdots \times B_d) = \{(X_1, \ldots, X_d) \in B_1 \times \cdots \times B_d\} = \{X_1 \in B_1\} \cap \cdots \cap \{X_d \in B_d\} \in \mathscr{F}$. □

Moreover, and more importantly for all the theory of stochastic processes, we can carry this to infinity. Let $(S_i, \mathscr{S}_i)$, $i = 1, 2, \ldots$, be measurable spaces. We wish to consider the space $S = \times_{i=1}^{\infty} S_i$ and endow it with a natural σ-field $\mathscr{S}$. To do so, consider sets of the form $B_1 \times B_2 \times \cdots$[b] where $B_i \in \mathscr{S}_i$ for all i, and where $B_j = S_j$ for all but finitely many indices j. Define $\mathscr{S} = \bigotimes_{i=1}^{\infty} \mathscr{S}_i$ to be the smallest σ field containing all these sets.

Lemma 3.6. *Suppose that for each $i \in \mathbb{N}$, $X_i : (\Omega, \mathscr{F}) \to (S_i, \mathscr{S}_i)$ is a random variable. Let $S = \bigtimes_{i=1}^{\infty} S_i$ and let $X = (X_1, X_2, \ldots) : (\Omega, \mathscr{F}) \to (S, \mathscr{S})$, where $\mathscr{S} = \bigotimes_{i=1}^{\infty} \mathscr{S}_i$ is the product-sigma field. Then X is a random variable.*

Proof Consider the π-system consisting of sets of the form $\cap_{i=1}^{\infty} B_i$ where $B_i \in \mathscr{S}_i$ for all i, and where $B_i = S_i$ for all but finitely many indices i. □

Now consider the set $\mathbb{R}^2$. The natural σ-field on it if $\mathscr{B} \otimes \mathscr{B}$, usually denoted by $\mathscr{B}(\mathbb{R}^2)$. Similarly for $\mathbb{R}^d$.

We should view $\sigma(X)$ as the essential information contained in X. We explain: If Ω is a set and $X : \Omega \to S$ a function then just by looking at the values of X we may not be able to distinguish elements of Ω, simply because X may take the same value on different points of Ω. In other words, unless X is one-to-one, the set $X(\Omega) = \{X(\omega), \omega \in \Omega\}$ contains “less information” than Ω. Take a trivial example: Let $\Omega = \{-N, -N+1, \ldots, N-1, N\}$, and let $X(i) = i^3$ for all $i \in \Omega$. Then $X(\Omega)$ contains precisely the same information as Ω. But if $Y(i) = i^2$ then $Y(\Omega)$ is less informative than Ω because the sign is lost. In some sense, we are not interested in the exact values of a random variable but only on whether the values can distinguish elements of Ω.

Take an abstract set Ω (for example think of Ω as a square in the 2-dimensional Euclidean plane), let $A \subset \Omega$ and consider the random variable $\mathbf{1}_A$. This takes only two values, 0 and 1, so the information conveyed by it is not that large. Indeed,

$$\sigma(\mathbf{1}_A) = \{\varnothing, A, A^c, \Omega\}$$

We may represent $\sigma(\mathbf{1}_A)$ as a partition of Ω, the partition consisting of A and its complement. Thus, $\mathbf{1}_A$ can only tell us whether we are on A or on its complement: very little information indeed. But we may as well use any function of $\mathbf{1}_A$ that does not compress it further, for example, $c \cdot \mathbf{1}_A$, where c is a nonzero constant.

EXERCISE 3. Consider two subsets A_1, A_2 of Ω and prove that the σ-field generated by $c_1\mathbf{1}_{A_1} + c_2\mathbf{1}_{A_2}$ is

$$\begin{aligned}\sigma(c_1\mathbf{1}_{A_1} + c_2\mathbf{1}_{A_2}) = \{&\varnothing,\ \Omega,\ A_1,\ A_2,\ A_1^c,\ A_2^c,\ A_1A_2,\ A_1A_2^c,\ A_1^cA_2,\ A_1^cA_2^c,\\ &A_1 \cup A_2,\ A_1 \cup A_2^c,\ A_1^c \cup A_2,\ A_1^c \cup A_2^c,\ A_1 \triangle A_2,\ (A_1 \triangle A_2)^c\}.\end{aligned}$$

[b] Each such set is called ‘finite-dimensional rectangle’.

It is more convenient to represent this σ-field by the partition of Ω induced by A_1, A_2, i.e. by the collection of disjoint sets

$$\{A_1A_2,\ A_1 \setminus A_2,\ A_2 \setminus A_1,\ (A_1A_2)^c\}.$$

EXERCISE 4. Any nonempty set in Exercise 3 can be obtained as union of elements of this partition.

EXERCISE 5. Given a partition $\mathscr{C} = \{C_1, \dots, C_n\}$ of Ω show that the σ-field generated by $\mathscr{C}$ consists of the empty set and all sets that can be obtained by taking unions of sets in $\mathscr{C}$. Assuming that none of the C_i is empty, this $\mathscr{C}$ contains exactly 2^n sets. Also show that any random variable that takes value b_i on C_i for each i generates a σ-field which is contained in $\sigma(\mathscr{C})$. Show that if the values b_i are distinct then $\sigma(X) = \sigma(\mathscr{C})$.

Unfortunately, this partition point of view does not carry over to big spaces. It works well for finite Ω or for general Ω but small σ-fields, but fails in richer situations. However, the intuition obtained remains, in some sense, valid. Let us explore this further.

If X, Y are random variables on a common measurable space $(\Omega, \mathscr{F})$ (and values in arbitrary sets) we say that Y IS MEASURABLE WITH RESPECT TO X if

$$\sigma(Y) \subset \sigma(X).$$

We often write this as

$$Y \in \sigma(X).$$

For example, if X is one random variable then X^2 is measurable with respect to X (but not vice-versa); if (X_1, X_2, X_3) are three random variables then (X_1, X_2) is measurable with respect to (X_1, X_2, X_3) and so is $(\cos(X_1 + X_2), -X_3 + \log|X_1|)$. If it appears that

> "measurability with respect to" means "(suitable) function of"

then this is because it really is so:

Lemma 3.7. *If $X : (\Omega, \mathscr{F}) \to (S, \mathscr{S})$, $Y : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B})$ are random variables, and if $Y \in \sigma(X)$, then there exists a random variable (measurable map) $H : (S, \mathscr{S}) \to (\mathbb{R}, \mathscr{B})$ such that $Y = H \circ X$.*

(The proof is deferred until the next page or so.) To understand why let us consider first a trivial situation: suppose that X takes only finitely many values; let $x_1, \dots, x_n$ be these (distinct) values.

EXERCISE 6. Show that if $X(\Omega) = \{x_1, \dots, x_n\}$ is a finite set then $\sigma(X)$ is generated by the partition $\big\{\{X = x_i\},\ i = 1, \dots, n\big\}$ of Ω.

EXERCISE 7. Using Exercise 5 show that any Y, measurable with respect to X, must be of the form

$$Y = \sum_{i=1}^{n} c_i \mathbf{1}(X = x_i).$$

We then see that what is claimed in Lemma 3.7 is correct, in this special situation with X being finitely-valued. Indeed, let $H(x) = \sum_{i=1}^{n} c_i \mathbf{1}(x = x_i)$ and, obviously, $Y = H \circ X$.

The general case requires an approximation result that says that any measurable random variable $X : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B})$ can be approximated by simple random variables. A SIMPLE RANDOM VARIABLE is a random variable with finitely many values.

Lemma 3.8. *Let $X : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B})$. Then there exists a sequence $X_1, X_2, \ldots$ of simple random variables such that $\lim_{n\to\infty} X_n(\omega) = X(\omega)$. If $X(\omega) \geq 0$, $\omega \in \Omega$, we can choose the sequence so that $0 \leq X_n(\omega) \leq X_{n+1}(\omega)$ for each n and ω.*

Proof First suppose $X(\omega) \geq 0$ for all $\omega \in \Omega$. Let $\lceil x \rceil$ denote the smallest integer k such that $k \geq x$. Define

$$\tau_n(x) := 2^{-n} \lceil 2^n x \rceil \wedge n. \tag{3.1}$$

$$X_n(\omega) := \tau_n(X(\omega)). \tag{3.2}$$

That $X_n(\omega) \to X(\omega)$ as $n \to \infty$ is obvious. Since $\lceil a \rceil \leq a$ we have $2\lceil a \rceil \leq \lceil 2a \rceil$ and so $2\lceil 2^n X \rceil \leq \lceil 2^{n+1} X \rceil$. Dividing both sides with 2^{n+1} we have $2^{-n}\lceil 2^n X \rceil \leq 2^{-n-1}\lceil 2^{n+1} X \rceil$ and so $X_n \leq X_{n+1}$. Next do not place any requirement on the sign of X but observe that

$$X = X^+ - X^-.$$

Then reduce to the previous case. □

Proof of Lemma 3.7 Suppose first that $Y = \mathbf{1}_A$ for some $A \in \mathscr{F}$ and that $Y \in \sigma(X)$. This means that $A \in \sigma(X)$. By definition of $\sigma(X)$, there is a $B \in \mathscr{B}$ such that $A = X^{-1}(B)$. Define $H(x) = \mathbf{1}(x \in B)$, $x \in \mathbb{R}$. Then $H(X(\omega)) = \mathbf{1}(X(\omega) \in B) = \mathbf{1}(\omega \in X^{-1}(B)) = \mathbf{1}(\omega \in A) = Y(\omega)$ and so $H \circ X = Y$, as required. Suppose next that Y is a simple random variable. Thus $Y = \sum_{i=1}^k c_i \mathbf{1}_{A_i}$ where $A_i \in \mathscr{F}$. By the previous case, there are $H_i : (S, \mathscr{S}) \to (\mathbb{R}, \mathscr{B})$, such that $\mathbf{1}_{A_i} = H_i(X)$. Define $H(x) = \sum_{i=1}^k c_i H_i(x)$. Then, clearly, $H(X) = Y$. Finally, suppose that $Y : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B})$ is a positive random variable. Then, by lemma 3.8, we can find simple $Y_n : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B})$, for all $n \in \mathbb{N}$, such that $Y = \sup_n Y_n$. By the case analysed above, we have, for each n, a measurable function H_n such that $Y_n = H_n \circ X$. Define $H = \sup_n H_n$. By Lemma 3.3, H is measurable. Also, $H \circ X = (\sup_n H_n) \circ X = \sup_n (H_n \circ X) = \sup_n Y_n = Y$, as required. In the general case, write $Y = Y^+ - Y^-$ and reduce to the above one. □

EXERCISE 8. Let X be one random variable in $\mathbb{R}$. Show that $\sigma(X^2)$ is a strict subset of $\sigma(X)$. But show that $\sigma(2^X) = \sigma(X)$.

Remark Lemma 3.7 is very important in enhancing our understanding of measurability. Measurability, in Probability Theory (as well as in other areas of Mathematics like Descriptive Set Theory) is not an abstract notion but conveys precisely the idea of "information" contained in a measurement. It should also be stressed that Y in Lemma 3.7 does not have to be restricted to take values in $\mathbb{R}$. It could very well be a random vector (values in $\mathbb{R}^d$) or, more generally, a random element of a fairly arbitrary space known as Polish space. We shall not enter into such advanced details here.

3.5 Law of a random variable

The reader will of course have noticed that

> the concept of "random variable" has nothing to do with randomness.

The adjective "random" is attached because, in addition to the above, we also consider some probability (measure) $\mathbf{P}$ on $(\Omega, \mathscr{F})$. As a mathematical definition, the concept of random variable is purely functional. So let us do that: let $\mathbf{P}$ be a probability on $(\Omega, \mathscr{F})$. The DISTRIBUTION or LAW of the random variable X is a probability $\mathbf{P}_X$ on $(S, \mathscr{S})$ which is induced, in the most natural fashion, by X:

$$\mathbf{P}_X(B) = \mathbf{P}(X^{-1}(B)), \quad B \in \mathscr{S}.$$

Note that $\mathbf{P}_X$ depends on two functions: the function $\mathbf{P}$ and the function X.

EXERCISE 9. Show that if $(\Omega, \mathscr{F}, \mathbf{P})$ is a probability space and $X : (\Omega, \mathscr{F}) \to (S, \mathscr{S})$ a random variable then $(S, \mathscr{S}, \mathbf{P}_X)$ is a probability space. Hence if X is a random vector then X_1 is a random variable.

EXERCISE 10. With the notation of Exercise 1, show that the law of $H \circ X$ as a random variable on the probability space $(\Omega, \mathscr{F}, \mathbf{P})$ is the same as the law of H as a random variable on $(S, \mathscr{S}, \mathbf{P}_X)$.

So the role of a random variable is to transform an abstract probability space into a concrete one.

In practise, one is often given[c] a probability measure $\mathbf{Q}$ on some $(S, \mathscr{S})$ and one may (or may not) want to construct a probability space $(\Omega, \mathscr{F}, \mathbf{P})$ and a random variable $X : (\Omega, \mathscr{F}) \to (S, \mathscr{S})$ such that $\mathbf{P}_X = \mathbf{P}$. In the absence of any further requirement we follow Occam's razor and consider the so-called CANONICAL CONSTRUCTION: take $\Omega = S$, $\mathscr{F} = \mathscr{S}$, $\mathbf{P} = \mathbf{Q}$ and let $X(\omega) \equiv \omega$. Then, obviously, $\mathbf{P}_X = \mathbf{Q}$. If this appears to be silly then this is because it is. But silly things are often quite useful. On the other hand, if other requirements are needed to be satisfied, choosing the "right" probability space and the "right" random variable is an art. The freedom in the choice of probability space can be roughly compared to the freedom of choice of an appropriate coördinate system in $\mathbb{R}^3$ when dealing, e.g. with the solution of a certain physical problem expressed in terms of a partial differential equation. For example, when we study the motion of fluid in a cylinder we may want to choose cylindrical coördinates.

3.6 Law of a discrete random variable

A DISCRETE RANDOM VARIABLE $X : (\Omega, \mathscr{F}) \to (S, \mathscr{S})$ is, by definition, one that takes countably many values. In other words, if $\mathbf{P}$ is a probability on $(\Omega, \mathscr{F})$ then X is discrete if and only if there is a countable set $D \in \mathscr{S}$ such that $\mathbf{P}(X \in D) = 1$. We also assume that D and all its subsets are members of $\mathscr{S}$. Hence the law $\mathbf{P}_X$ of X is a probability on D. We know that a probability on a countable set D can be defined by defining its values on singletons. These values form the so-called (in baby probability talk) PROBABILITY MASS FUNCTION (see $\boxed{1.1.5}$) Thus, the probability mass function is

$$p(x) = \mathbf{P}_X\{x\} = \mathbf{P}(X = x), \quad x \in D.$$

Clearly, if $B \subset D$ then

$$\mathbf{P}_X(B) = \mathbf{P}_X\Big(\bigcup_{x \in B} \{x\}\Big) = \sum_{x \in B} p(x).$$

So p is sufficient for computing $\mathbf{P}_X$.

Example 3.1. A box contains n socks labelled 1 through n. Pick two socks at random and let X be the pair of their labels. Then X is a discrete random variable taking values (i, j) where $1 \le i, j \le n$, $i \ne j$. The probability mass function is

$$p(i, j) = 1/n(n-1).$$

But what is the probability space? First, we should realise that we do not necessarily need to consider it. Second, we should realise that the article "the" is wrong, for there are many choices for a probability space. Having said that, a reasonable choice for Ω is the set of permutations

[c] Actually, one is seldom given anything. Either one derives something from some basic principles/requirements or one performs an experiment whereby measurements are collected and a probability measure is stipulated. The latter is the subject of Statistics

ω of $\{1, \dots, n\}$, i.e. $\omega = (\omega_1, \dots, \omega_n)$ where all the ω_i take values in $\{1, \dots, n\}$ and are distinct. The σ-field $\mathscr{F}$ should be rich enough to contain all singletons, i.e. $\mathscr{F} = 2^\Omega$. The probability $\mathbf{P}$ should be defined as $\mathbf{P}(\{\omega\}) = 1/n!$ for all ω. To be honest we should check that $p(i,j) = \mathbf{P}\{\omega \in \Omega : \omega_1 = i, \omega_2 = j\}$. We have

$$\mathbf{P}\{\omega \in \Omega : \omega_1 = i, \omega_2 = j\} = \frac{1}{n!} \sharp\{\omega \in \Omega : \omega_1 = i, \omega_2 = j\} = \frac{(n-1)!}{n!} = \frac{1}{n(n-1)},$$

as required. So $(\Omega, \mathscr{F}, \mathbf{P})$ is a probability space, X is a random variable (check this!), $\mathbf{P}_X$ is its law, and p is its probability mass function.

Example 3.2. Consider the experiment of picking k integers at random (with replacement) from the set $\{1, \dots, n\}$. Our $(\Omega, \mathscr{F})$ here can be taken to be

$$\Omega = \{1, \dots, n\}^k, \quad \mathscr{F} = 2^\Omega,$$

and the probability corresponding to the experiment should be defined as

$$\mathbf{P}\{(\omega_1, \dots, \omega_k)\} = \frac{1}{n^k}.$$

Note here that $\{(\omega_1, \dots, \omega_k)\}$ is a set containing a single point, namely the ordered k-tuple $\omega = (\omega_1, \dots, \omega_k)$. For more general sets B we take, as required, $\mathbf{P}(B) = \sum_{\omega \in B} \mathbf{P}\{\omega\} = \frac{\sharp B}{n^k}$. Define a random variable representing the maximum number picked:

$$X(\omega) = \max(\omega_1, \dots, \omega_k).$$

Clearly, X takes values in $\{1, \dots, n\}$ and trivially, X is a measurable function from $(\Omega, \mathscr{F})$ to $(\{1, \dots, n\}, 2^{\{1, \dots, n\}})$. (The inverse image of a subset of $\{1, \dots, n\}$ under X is, of course, a subset of Ω, i.e. an element of $\mathscr{F}$.) Let us compute the probability mass function for X. For $x \in \{1, \dots, n\}$ we have

$$\begin{aligned}
p(x) &= \mathbf{P}(X^{-1}(x)) = \mathbf{P}(X = x) = \mathbf{P}\{\omega \in \Omega : \ \max(\omega_1, \dots, \omega_k) = x\} \\
&= \mathbf{P}\{\omega \in \Omega : \ \max(\omega_1, \dots, \omega_k) \le x \text{ but } \max(\omega_1, \dots, \omega_k) \not\le x - 1\} \\
&= \mathbf{P}\{\omega \in \Omega : \ \max(\omega_1, \dots, \omega_k) \le x\} - \mathbf{P}\{\omega \in \Omega : \ \max(\omega_1, \dots, \omega_k) \le x - 1\} \\
&= \frac{1}{n^k}\sharp\{\omega \in \Omega : \ \omega_1 \le x, \dots, \omega_k \le x\} - \frac{1}{n^k}\sharp\{\omega \in \Omega : \ \omega_1 \le x-1, \dots, \omega_k \le x-1\} \\
&= \frac{x^k - (x-1)^k}{n^k}.
\end{aligned}$$

EXERCISE 11. Suppose that the random variable X takes n distinct values (i.e. $X(\Omega)$ is a set with n elements). Show that $\sigma(X)$ has 2^{2^n} elements and describe (give a procedure for describing) them.

3.7 Tossing coins

Consider the experiment of flipping fair coins independently (see $\boxed{2.2}$). A coin takes 2 values, heads or tails with probability 1/2 each. Label heads by 1 and tails by 0. Let ω_1 be outcome of the first toss, ω_2 that of the second, etc. We stipulate that:

A coin flipped n times yields a string $(\omega_1, \dots, \omega_n)$ of n 0s or 1s with probability $1/2^n$ each.

For example $(0,1,0)$ should have probability $1/8$, while $(1,1,0,0,1)$ should have probability $1/32$. In other words, we are *trying* to define a probability $\mathbf{P}$ by saying that the $\mathbf{P}$ of an event that specifies the first n outcomes completely should be $1/2^n$, for each $n \in \mathbb{N}$. Consider our desire to find out when we will get our first 1 (head), i.e. let

$$T = \inf\{n \in \mathbb{N} : \ \omega_n = 1\}.$$

This *appears* to be a random variable. But on which measurable space? We cannot pick Ω to be $\{0,1\}^n$ for some finite n, because we have no bound on T. We can, however, pick

$$\Omega = \{0,1\}^{\mathbb{N}},$$

i.e. the set of all infinite-length sequences $\omega = (\omega_1, \omega_2, \ldots)$, where $\omega_n = 0$ or 1 for all n. This is our COIN-FLIP SPACE.

EXERCISE 12. Show that the coin-flip space $\{0,1\}^{\mathbb{N}}$ is uncountable.

To put what we said in some notation, let

$$R(i_1, \ldots, i_n) := \{\omega \in \Omega : \ \omega_1 = i_1, \ldots, \omega_n = i_n\}.$$

To this set we should assign probability 2^{-n}:

$$\mathbf{prob}(R(i_1, \ldots, i_n)) = 2^{-n}. \tag{3.3}$$

We use the symbol **prob** rather than $\mathbf{P}$ because we have not yet verified that this is a probability. Let $\mathscr{R}$ be the class of such sets (together with the empty set):

$$\mathscr{R} := \{\varnothing\} \cup \{R(i_1, \ldots, i_n) : \ i_1, \ldots, i_n \in \{0,1\}, \ n \in \mathbb{N}\}.$$

Formula (3.3) defines the function

$$\mathbf{prob} : \mathscr{R} \to \mathbb{R}.$$

It is by no means true that $\mathscr{R}$ contains all subsets of Ω (far from that):

EXERCISE 13. Show that

$$R(1) \cup R(0,1) \cup R(0,0,1) \cup \cdots \notin \mathscr{R}.$$

Show that this set is the set $\{\omega \in \Omega : \ T(\omega) < \infty\}$. Show that the intersection of two elements of $\mathscr{R}$ is in $\mathscr{R}$, but not the union.

EXERCISE 14. Second, the complement of any set in $\mathscr{R}$ is written as a finite disjoint union of sets in $\mathscr{R}$. (For example $R(0,1)^c = \{\omega : \omega_1 = 1\} \cup \{\omega : \omega_2 = 0\} = R(1) \cup \{\omega : \omega_2 = 0\} = R(1) \cup R(1,0) \cup R(0,0)$.)

What kind of properties does **prob** have? We would like **prob** to be, at least, additive, i.e. to satisfy $\mathbf{prob}(G_1 \cup \cdots \cup G_k) = \mathbf{prob}(G_1) + \cdots + \mathbf{prob}(G_k)$ if $G_1, \ldots G_k$ are mutually disjoint elements of $\mathscr{R}$. There is an issue here: as we saw, the union of elements of $\mathscr{R}$ may not be in $\mathscr{R}$. So this last additivity property is completely meaningless for **prob**. We remedy this by enlarging $\mathscr{R}$:

$$\mathscr{C} := \{G_1 \cup \cdots \cup G_k : \ G_1, \ldots, G_k \text{ are mutually disjoint elements of } \mathscr{R}, \ k \in \mathbb{N}\},$$

and by extending **prob** to $\mathscr{C}$ in the most natural way:

$$\mathbf{P}_0(G_1 \cup \cdots \cup G_k) := \sum_{j=1}^{k} \mathbf{prob}(G_j).$$

Lemma 3.9. *(i) $\mathscr{C}$ is a field. (ii) $\mathbf{P}_0$ is well-defined and additive over $\mathscr{C}$.*

Proof (i) Let $A, B \in \mathscr{C}$. Then $A = \cup_{i=1}^n G_i$, $B = \cup_{j=1}^m H_j$, where the G_i are mutually disjoint (m.d.) elements of $\mathscr{C}$, and so are the H_j. We have

$$AB = \cup_{i,j} G_i H_j$$

and, clearly, the $G_i H_j$ are m.d. elements of $\mathscr{R}$; so $AB \in \mathscr{C}$. Also,

$$A^c = \cap_j G_i^c,$$

each G_i^c is an element of $\mathscr{C}$ (Exercise 14), and because we just proved that $\mathscr{C}$ is closed under intersections, we have $A^c \in \mathscr{C}$. So $\mathscr{C}$ is a field.
(ii) To show that $\mathbf{P}_0$ is well-defined means to show that if we can write the same $A \in \mathscr{C}$ in two different ways, say $A = \cup_{i=1}^n G_i = \cup_{j=1}^m H_j$ where both the G_i and the H_j are m.d. elements of $\mathscr{R}$, we have $\sum_i \mathbf{prob}(G_i) = \sum_j \mathbf{prob}(H_j)$. But notice that $G_i = \cup_j G_i H_j$ and so $\mathbf{prob}(G_i) = \sum_j \mathbf{prob}(G_i H_j)$, and, similarly, $\mathbf{prob}(H_j) = \sum_i \mathbf{prob}(G_i H_j)$. This proves the claim that $\mathbf{P}_0$ is well-defined on $\mathscr{C}$. Now, to show additivity, suppose that $A_1, \ldots, A_n$ are m.d. elements of $\mathscr{C}$. We must show that $\mathbf{P}_0(\cup_i) = \sum_i \mathbf{P}_0(A_i)$. Now, to show additivity, suppose that $A_1, \ldots, A_n$ are m.d. elements of $\mathscr{C}$. We must show that $\mathbf{P}_0(\cup_i A_i) = \sum_i \mathbf{P}_0(A_i)$. But each A_i is itself a finite union of m.d. elements of $\mathscr{R}$, say $A_i = \cup_j H_{ij}$, and, by the definition of $\mathbf{P}_0$, we have $\mathbf{P}_0(A_i) = \sum_j \mathbf{prob}(H_{ij})$. So $\sum_i \mathbf{P}_0(A_i) = \sum_i \sum_j \mathbf{prob}(H_{ij})$. On the other hand, $\cup_i A_i = \cup_{i,j} H_{ij}$ and the H_{ij} are m.d. elements of $\mathscr{R}$. Hence, by the definition of $\mathbf{P}_0$ again, we have $\mathbf{P}_0(\cup_i A_i) = \sum_i \sum_j \mathbf{prob}(H_{ij})$. □

To put in words, what we have shown so far is that, starting merely from the intuitive definition (3.3), we can compute probabilities of sets that can be expressed using finitely many outcomes, i.e. that involve ω_i up to some finite index. Mathematically, Lemma 3.9 has enabled us to define a function

$$\mathbf{P}_0 : \mathscr{C} \to \mathbb{R}$$

in such a way that

$$\mathbf{P}_0(G) = \mathbf{prob}(G), \quad \text{if } G \in \mathscr{R}$$

and

$$\mathbf{P}_0(G_1 \cup \cdots \cup G_k) = \mathbf{P}_0(G_1) + \cdots + \mathbf{P}_0(G_k), \quad \text{if } G_1, \ldots G_k \text{ are mutually disjoint elements of } \mathscr{R}$$

(a property which, as remarked, is meaningless for **prob**).

But there are many sets, whose probability we would like to know, but which do not belong to $\mathscr{C}$.

Example 3.3. The following (useful) sets do not belong to $\mathscr{C}$:
(i) $A_1 := \{\omega : \; T(\omega) < \infty\}$ (i.e. the event that a head will occur at some point). Indeed, this equals $\cup_{n=1}^\infty \{\omega \in \Omega : \; \omega_n = 1\}$ and does not belong to $\mathscr{C}$ because it involves infinitely many coordinates.
(ii) $A_2 := \{\omega : \omega_n = 1 \text{ for infinitely many } n\}$. This, obviously, does not belong to $\mathscr{C}$. Formally the set equals

$$\{\omega : \; \forall n \; \exists m \geq n \; \omega_m = 1\} = \bigcap_n \bigcup_{m \geq n} \{\omega : \; \omega_m = 1\} = \bigcap_n \bigcup_{m \geq n} \bigcup_{i_1, \ldots, i_m \in \{0,1\}} R(i_1, \ldots, i_{m-1}, 1).$$

And so it is explicitly seen that we need infinitely many sets from $\mathscr{R}$ to construct A_2.
Although we feel that both A_1 and A_2 should be assigned probability 1, we *cannot* write $\mathbf{P}_0(A_1) = 1$, neither $\mathbf{P}_0(A_2) = 1$, simply because neither of the two sets belongs to the domain of the function $\mathbf{P}_0$.

So our desire to deal with these sets (and not only!) forces us to carry the story further.

We saw (Lemma 3.9 that $\mathbf{P}_0$ is additive on $\mathscr{C}$. But more is true:

Lemma 3.10. $\mathbf{P}_0$ *is countably additive over* $\mathscr{C}$*, that is, if* G_n *are mutually disjoint elements of* $\mathscr{C}$ *and if* $\cup_n G_n \in \mathscr{C}$ *then* $\mathbf{P}_0(\cup_n G_n) = \sum_n \mathbf{P}_0(G_n)$*.*

Proof By Lemma [CA], Appendix B, it suffices to show that if G_n is a decreasing sequence of elements of $\mathscr{C}$ such that $\cap_n G_n = \varnothing$ then $\mathbf{P}_0(G_n) \to 0$. − To be continued − □

We next consider the class $D(\mathscr{C})$ of sets that are limits of sequences of sets in $\mathscr{C}$. Lemma [FIRST EXTENSION], Appendix B, shows that we can define $\mathbf{P}_1$ on $D(\mathscr{C})$ such that $\mathbf{P}_1$ is countably additive and agrees with $\mathbf{P}_0$ on $\mathscr{C}$.

We then consider the class $D^2(\mathscr{C}) = D(D(\mathscr{C}))$ of sets that are limits of sequences of sets in $D(\mathscr{C})$. Just as above, we can define $\mathbf{P}_2$ on $D^2(\mathscr{C})$ such that $\mathbf{P}_2$ is countably additive and agrees with $\mathbf{P}_1$ on $D(\mathscr{C})$.

One would suspect that by continuing, in this manner, by induction, i.e. by taking sets which are limits of limits of limits, ad infinitum, we would exhaust the class

$$\mathscr{F} = \sigma\text{-field generated by } \mathscr{C}. \tag{3.4}$$

(It is not difficult to see that, also,

$$\mathscr{F} = \sigma\text{-field generated by } \mathscr{R}.) \tag{3.5}$$

Unfortunately, this does not work. The way out is to do as in Proposition [SANDWICH], Appendix B. This proposition actually proves the following:

Theorem 3.1. *Let* $\Omega = \{0,1\}^{\mathbb{N}}$*. Let* $\mathscr{R}$ *be the class of sets of the form* $R(i_1, \ldots, i_n) := \{\omega \in \Omega : \ \omega_1 = i_1, \ldots, \omega_n = i_n\}$*. Define* $\mathbf{prob}(R(i_1, \ldots, i_n)) = 2^{-n}$*. Let* $\mathscr{F}$ *be the* σ*-field generated by* $\mathscr{R}$*. Then there exists a UNIQUE probability* $\mathbf{P}$ *on* $(\Omega, \mathscr{F})$ *such that* $\mathbf{P}(G) = \mathbf{prob}(G)$ *for every* $G \in \mathscr{R}$*.*

Proof See proof of Proposition [SANDWICH], Appendix B. To prove uniqueness, assume that there is another probability $\widetilde{\mathbf{P}}$ on $(\Omega, \mathscr{F})$ that agrees with **prob** on $\mathscr{R}$. Then $\widetilde{\mathbf{P}}$ agrees with **prob** (and hence with $\mathbf{P}$ on $\mathscr{C}$. Let us define the class common domain of $\mathbf{P}$, $\widetilde{\mathbf{P}}$:

$$\mathscr{D} := \{A \in 2^{\Omega} : \ \mathbf{P}(A) = \widetilde{\mathbf{P}}(A)\}.$$

We have

$$\mathscr{C} \subset \mathscr{D}.$$

Therefore, of $\lambda(\mathscr{C})$ (resp. $\lambda(\mathscr{D})$) is the smallest λ-system containing $\mathscr{C}$ (resp. $\mathscr{D}$) we have

$$\lambda(\mathscr{C}) \subset \lambda(\mathscr{D}).$$

But it is easy to see that $\mathscr{D}$ is itself closed under proper differences and increasing limits. Therefore $\lambda(\mathscr{D}) = \mathscr{D}$. On the other hand, by Lemma [Sierpiński-Dynkin], Appendix B, we have $\lambda(\mathscr{C}) = \sigma(\mathscr{C})$. So $\mathscr{F} := \sigma(\mathscr{C})$ is contained in $\mathscr{D}$. In other words, the two probabilities agree on $\mathscr{F}$. □

This theorem tells us that

$$\boxed{(\{0,1\}^{\mathbb{N}}, \mathscr{F}, \mathbf{P}) \text{ is a probability space.}}$$

So we can use all the good things we know about a probability space. For example, $T(\omega) = \inf\{n \in \mathbb{N} : \omega_n = 1\}$ defines a measurable function $T : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B})$ and, since $\{T \le n\}$ are increasing sets with union equal to the set $\{T < \infty\}$, we have $\mathbf{P}(T < \infty) = \lim_{n\to\infty} \mathbf{P}(T \le n)$. But $\{T > n\} = R(0, \dots, 0)$ (n 0's inside the parenthesis) and so $\mathbf{P}(T > n) = \mathbf{prob}(R(0, \dots, 0)) = 2^{-n}$. Thus $\mathbf{P}(T < \infty) = \lim_n (1 - 2^{-n}) = 1$.

EXERCISE 15. Consider the space $\Omega = \{0,1\}^{\mathbb{N}}$ and the class of sets $\mathscr{R}$ defined in the tossing coins section. (i) Show that $\mathscr{R}$ is a countable set, i.e. there is a one-to-one correspondence between $\mathscr{R}$ and $\mathbb{N}$. (ii) Show that $\mathscr{C}$ is the field generated by $\mathscr{R}$ and is also countable. (iii) Consider $\mathscr{F}$, the σ-field generated by $\mathscr{R}$ (and also by $\mathscr{C}$) and show that it is uncountable; more specifically, show that there is a one-to-one correspondence between $\mathscr{F}$ and $2^{\mathbb{N}}$. (iv) Consider the class 2^{Ω} of all subsets of Ω and show that there is a one-to-one correspondence between 2^{Ω} and $2^{2^{\mathbb{N}}}$ (the class of subsets of the set of subsets of $\mathbb{N}$.) It is known [3] that the cardinality of the set of subsets of a set is strictly larger than the cardinality of the set itself (Schroeder-Bernstein theorem). Using this, conclude that there 2^{Ω} is strictly larger than $\mathscr{F}$. In fact MUCH larger: In some sense, most subsets of Ω are outside $\mathscr{F}$! This means that most sets of sequences of coin tosses cannot be expressed using countably many operations on elementary sets (i.e. sets from $\mathscr{R}$). The CURIOUS thing though is that we know very few of these sets: although the majority of them are outside $\mathscr{F}$, somehow, it is very heard to come up with an example.[d]

3.8 Uniform random variable

We will now construct our first non-discrete random variable. Consider the probability space $(\Omega = \{0,1\}^{\mathbb{N}}, \mathscr{F}, \mathbf{P})$ of the coin tossing experiment. Define

$$U(\omega) = \frac{\omega_1}{2} + \frac{\omega_2}{2^2} + \frac{\omega_3}{2^3} + \cdots = \sum_{k=1}^{\infty} \frac{\omega_k}{2^k}. \tag{3.6}$$

Let $U_n(\omega) := \sum_{k=1}^{n} \frac{\omega_k}{2^k}$: these are simple random variables. Since $U(\omega) = \lim_{n\to\infty} U_n(\omega)$, we have, by Lemma 3.3, that $U : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B})$ is measurable (i.e. a random variable). It is ONE random variable because it takes values in $\mathbb{R}$. More specifically, observe that

$$0 \le U(\omega) \le \sum_{k=1}^{\infty} \frac{1}{2^k} = 1.$$

So $U(\Omega) \subset [0,1]$. Our first observation is

Lemma 3.11. *For any integer $0 \le a < 2^n$ and any $n \in \mathbb{N}$, unless $\omega_i = 1$ for all $i > n$,*

$$U \ge \frac{a}{2^n} \iff U_n \ge \frac{a}{2^n}.$$

Proof Note that $2^n U_n$ is an integer. If it is not the case that $\omega_i = 1$ for all $i > n$ we have $2^n U - 2^n U_n < 1$, i.e. the integer part of $2^n U$ equals $2^n U_n$ and hence $2^n U \ge a$ iff $2^n U_n \ge a$. □

Next, we compute the following probabilities:

Lemma 3.12. *For any integer $0 \le a < 2^n$ and any $n \in \mathbb{N}$, we have*

$$\mathbf{P}(U \le a/2^n) = a/2^n.$$

[d] In some sense, this exercise is a baby version of Gödel's theorem which says, naively speaking, that most theorems cannot be proved. Except that we hardly ever encounter them.

Proof Using Lemma 3.11 we find that

$$\mathbf{P}(U < a/2^n) = \mathbf{P}(U_n < a/2^n).$$

We use induction on n. To start, let $n = 1$. Observe that $\mathbf{P}(U_1 < 1/2) = \mathbf{P}(\omega_1 = 0) = 1/2$, as required. Now assume the statement to be true up to index $n - 1$. Write

$$\begin{aligned} U_n(\omega) &= \frac{\omega_1}{2} + \frac{1}{2} Y_{n-1}(\omega), \\ Y_{n-1}(\omega) &= \frac{\omega_2}{2} + \frac{\omega_3}{2^2} + \cdots + \frac{\omega_{n-1}}{2^{n-1}}. \end{aligned}$$

Therefore

$$\begin{aligned} \mathbf{P}(U_n < a/2^n) &= \mathbf{P}(\omega_1 = 0,\ Y_{n-1} < a/2^{n-1}) + \mathbf{P}(\omega_1 = 1,\ Y_{n-1} < a/2^{n-1} - 1) \\ &= \tfrac{1}{2}\mathbf{P}(Y_{n-1} < a/2^{n-1}) + \tfrac{1}{2}\mathbf{P}(Y_{n-1} < a/2^{n-1} - 1). \end{aligned}$$

By the induction hypothesis, $\mathbf{P}(Y_{n-1} < c/2^{n-1}) = c/2^{n-1}$ for all $c < 2^{n-1}$ (why?). If $a \le 2^{n-1}$, the second term in the display is 0 and the first equals $(1/2) \times (a/2^{n-1}) = a/2^n$, as needed. If $2^{n-1} < a < 2^n$, the first term equals $1/2$ and the second equals $(1/2) \times (a/2^{n-1} - 1)$; the sum gives again $a/2^n$. □

Finally, we have

Lemma 3.13. *For any real number $x \in [0, 1]$,*

$$\mathbf{P}(U < x) = \mathbf{P}(U \le x) = x.$$

Proof Let $[y]$ denote the largest integer not exceeding the real number y. Define, recursively,

$$\begin{aligned} z_1 &= x, \quad \xi_1 = [2z], \\ z_{n+1} &= z_n - \frac{\xi_n}{2^n}, \quad \xi_{n+1} = [2^{n+1} z_{n+1}]. \end{aligned}$$

We have $2^n z_{n+1} = 2^n z_n - \xi_n = 2^n z_n - [2^n z_n] \le 1$, so $z_{n+1} \le 1/2^n$ for all n. Therefore the sequence z_n converges to 0 and the numbers ξ_n are either 0 or 1. We also have

$$x_n := \frac{\xi_1}{2} + \frac{\xi_2}{2^2} + \cdots + \frac{\xi_n}{2^n} = (z_1 - z_2) + (z_2 - z_3) + \cdots + (z_n - z_{n+1}) = x - z_{n+1},$$

and so $x_n \to x$, as $n \to \infty$. Also, x_n is of the form $a/2^n$, where a is an integer. So $\mathbf{P}(U \le x_n) = x_n$. Since x_n actually increases to x, we have $\cup_{n=1}^{\infty}\{U \le x_n\} = \{U < x\}$, and so taking limits we get $\mathbf{P}(U < x) = x$, for all $x \le 1$. This also implies $\mathbf{P}(U \le x + 1/n) = x + 1/n$ for all n sufficiently large. But $\cap_n \{U \le x + 1/n\} = \{U \le x\}$ and so, taking limits, $\mathbf{P}(U \le x) = x$. □

Corollary 3.2.

$$\mathbf{P}(U = x) = 0, \ \textit{for all } x \in [0, 1].$$

So the random variable U takes no specific value with positive probability. Yet, U exists (we constructed it). This is curious, but we can (must) get used to it. Any real random variable with this property is called CONTINUOUS RANDOM VARIABLE. In books, U has a name: it is called UNIFORM RANDOM VARIABLE ON THE INTERVAL $[0, 1]$.

Corollary 3.3. *Let $\mathbb{Q}$ be the set of all rational numbers in $[0, 1]$. Then*

$$\mathbf{P}(U \in \mathbb{Q}) = 0.$$

Proof The set $\mathbb{Q}$ is enumerable (has countably many points). Hence

$$\mathbf{P}(U \in \mathbb{Q}) = \mathbf{P}(\cup_{q \in \mathbb{Q}}\{U = q\}) = \sum_{q \in \mathbb{Q}} \mathbf{P}(U = q) = 0.$$

□

3.9 Probability, revisited

Recall, once more, that a probability $\mathbf{P}$ defined on a σ-field $\mathscr{F}$ of subsets of a set Ω is a function $\mathbf{P} : \mathscr{F} \to \mathbb{R}$ such that (i) $\mathbf{P}(\Omega) = 1$, and (ii) $\mathbf{P}(\cup_n A_n) = \sum_n \mathbf{P}(A_n)$, if the A_n are mutually disjoint elements of $\mathscr{F}$.

If Ω is a countable set then we know that it suffices to define $\mathbf{P}$ on singletons, i.e. knowledge of $\mathbf{P}\{\omega\}$ for all $\omega \in \Omega$ implies knowledge of $\mathbf{P}(A)$ for all $A \in \mathscr{F}$. Indeed,

$$\mathbf{P}(A) = \sum_{\omega \in A} \mathbf{P}\{\omega\}.$$

We saw, in the previous section, the necessity to define $\mathbf{P}$ on sets which are not countable. Can't we define $\mathbf{P}\{\omega\}$ for all ω in this case? The argument is simple. Suppose we did do so. Let $\Omega_+ := \{\omega \in \Omega : \ \mathbf{P}\{\omega\} > 0\}$, and, for each $n \in \mathbb{N}$, $\Omega_n = \{\omega \in \Omega : \ \mathbf{P}\{\omega\} \geq 1/n\}$. This is a finite set with at most n points. Observe that $\cup_n \Omega_n = \Omega_+$. But this tells us that Ω_+ is a countable set. So this poses a "philosophical" problem: If we want to define a random variable that is uniformly distributed on the uncountable set $[0, 1]$ by defining probabilities of singletons, then we'd have to pick a countable subset of $[0, 1]$ and define probabilities of points there. But which one to pick? We cannot give preference to any particular countable subset.

The way out is not to define probabilities on single points, but, rather, on sets, as we did in the previous section. We then do not give preference to any particular countable subset, but we lose something "intuitive": we assign zero probability to every individual point. This is what needs to be done, both for physical and mathematical reasons, and this is what shall be done hereafter: We will be defining probabilities on sets.

But another burden appears now: whereas, in the countable case, we were free to define probabilities of points at will (as long as they summed up to something finite), in the general case we cannot define probabilities of sets at will: we must respect the countable additivity property (ii) above. Since there are far too many sets, we will try to see how to define $\mathbf{P}$ on a class of selected few of them. So read the next section.

3.10 Distribution functions

Consider a real random variable $X : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B})$ and put a probability $\mathbf{P}$ on Ω. Recall the definition of the law $\mathbf{P}_X$ of X:

$$\mathbf{P}_X(B) = \mathbf{P}(X \in B), \quad B \in \mathscr{B}.$$

Up until the middle of the 19th century, a function was an object that was given by a formula. Nowadays, of course, a function is more abstract: we certainly need no formulae to comprehend what a function is. However, let us take the old-fashioned point of view. Clearly, $\mathbf{P}_X$ is a function whose arguments are sets (elements of $\mathscr{B}$). These are a lot of arguments! And, as we know, we don't need all of them because, for example, if $B = B_1 \cup B_2$ and $B_1 \cap B_2 = \varnothing$, then $\mathbf{P}_X(B) = \mathbf{P}_X(B_1) + \mathbf{P}_X(B_2)$. To put it pedantically, we can't define $\mathbf{P}_X(B)$ anyway we like. We must make sure that it is consistent (countable additivity must hold).

> So the question is: can we throw away some (many) of the values $\mathbf{P}(B)$ without losing information?

The answer is yes. It suffices to know $\mathbf{P}_X(B)$ only for B of specific form. There are many choices. One choice is to take

$$B = (-\infty, x],$$

where x ranges from $-\infty$ to ∞ and consider only (see $\boxed{\text{Example 1.5}}$)

$$F(x) := \mathbf{P}_X(-\infty, x], \quad x \in \mathbb{R}. \tag{3.7}$$

So I claim that:

Lemma 3.14. *Knowledge of $\mathbf{P}_X$ on this class of sets only (semi-infinite intervals) implies knowledge of $\mathbf{P}_X$ on the whole of $\mathscr{B}$.*

This is a big claim. But we shall prove it by construction. First, let us see the properties of F.

Lemma 3.15. *(i) $x_1 < x_2 \Rightarrow F(x_1) \le F(x_2)$, (ii) $\lim_{x\to-\infty} F(x) = 0$, (iii) $\lim_{x\to+\infty} F(x) = 1$, (iv) $\lim_{n\to\infty} F(x + 1/n) = F(x)$.*

Proof (i) If $x_1 < x_2$ we have $(-\infty, x_1] \subset (-\infty, x_2]$ and so $\mathbf{P}_X(-\infty, x_1] \le \mathbf{P}_X(-\infty, x_2]$. (ii)&(iii) Since F is monotonic with values between 0 and 1 it has a limit both as x tends to $-\infty$ and $+\infty$. We have $\cap_{n\in\mathbb{Z}}(-\infty, n] = \varnothing$, so $\mathbf{P}_X(-\infty, n] \to 0$ as $n \to -\infty$. Similarly, $\cup_{n\in\mathbb{Z}}(-\infty, n] = \mathbb{R}$, so $\mathbf{P}_X(-\infty, n] \to \mathbf{P}_X(\mathbb{R}) = 1$ as $n \to +\infty$. (iv) $\cap_n(-\infty, x + 1/n] = (-\infty, x]$. □

Remarks:
1. Notice that $F(x - 1/n)$ does NOT necessarily converge to $F(x)$ because $\cup_n(-\infty, x - 1/n] = (-\infty, x)$, and so $\lim_{n\to\infty} F(x - 1/n) = \mathbf{P}_X(-\infty, x)$ which will be different from $\mathbf{P}_X(-\infty, x]$ if $\mathbf{P}_X\{x\} > 0$.
2. There is no good reason that we chose this class of sets, other than people have been using it by convention. For instance, we could have chosen open semi-infinite interval $(-\infty, x)$, in which case (iv) of Lemma 3.15 would be replaced by $\lim_{n\to\infty} F(x - 1/n) = F(x)$.

Proof of Lemma 3.14 Consider the coin-flip $\Omega = \{0,1\}^{\mathbb{N}}$. $U(\omega) = \sum_{n=1}^{\infty} \frac{\omega_n}{2^n}$. Then U is a random variable and $\mathbf{P}$ is, indeed, a probability on $(\Omega, \mathscr{F})$, where $\mathscr{F}$ is defined through (3.4) or (3.5). We also showed that $\mathbf{P}(U \le u) = u$ for all $u \in [0,1]$. Suppose now that F, as in (3.7), is given. Define

$$F^{-1}(u) := \sup\{x \in \mathbb{R} : \ F(x) \le u\}. \tag{3.8}$$

By Lemma 3.3, $F^{-1}(U)$ is a random variable. In other words, $F^{-1}\circ U : (\Omega, \mathscr{F}) \to (\mathscr{R}, \mathscr{B})$ is measurable. Suppose that, for some u, t, we have $F^{-1}(u) \le t$. This means that t is an upper bound of the set $\{x \in \mathbb{R} : \ F(x) \le u\}$. Hence for all $x \in \mathbb{R}$, if $F(x) \le u$ then $x \le t$. Equivalently, for all $x \in \mathbb{R}$, if $x > t$ then $F(x) > u$. Equivalently, for all $\varepsilon > 0$, $F(t + \varepsilon) > u$. Equivalently, for all $n \in \mathbb{N}$, $F(t + 1/n) > u$. Thus,

$$\{\omega \in \Omega : \ F^{-1}(U(\omega)) \le t\} = \{\omega \in \Omega : \ \forall n \in \mathbb{N} \ F(t + 1/n) > U(\omega)\}.$$

Hence the probabilities of the two sets (notice they both belong to $\mathscr{F}$) are the same:

$$\mathbf{P}(F^{-1}(U) \le t) = \mathbf{P}(\forall n \in \mathbb{N} \ F(t + 1/n) > U).$$

By the fact that $\mathbf{P}(A_n) \to \mathbf{P}(A)$ if A_n is a decreasing sequence of sets with intersection A, we have

$$\mathbf{P}(\forall n \in \mathbb{N} \ F(t + 1/n) > U) = \lim_{n\to\infty} \mathbf{P}(U < F(t + 1/n)).$$

By Lemma 3.13, and the right-continuity of F (Lemma 3.15),

$$\lim_{n\to\infty} \mathbf{P}(F(t + 1/n) > U) = \lim_{n\to\infty} F(t + 1/n) = F(t).$$

Thus the random variable $F^{-1}\circ U$ has the same law as X. Hence, for all $B \in \mathscr{B}$,

$$\mathbf{P}_X(B) = \mathbf{P}(F^{-1}(U) \in B) = \mathbf{P}\{\omega \in \{0,1\}^{\mathbb{N}} : \ F^{-1}(U(\omega)) \in B\}$$

So, knowledge of $\mathbf{P}_X$ on sets of the form $(-\infty, x]$ implies knowledge of $\mathbf{P}_X(B)$ for all $B \in \mathscr{B}$ by the explicit formula of the last display. □

Definition 3.1. A function $F : \mathbb{R} \to \mathbb{R}$ is called distribution function iff (i) $x_1 < x_2 \Rightarrow F(x_1) \leq F(x_2)$, (ii) $\lim_{x \to -\infty} F(x) = 0$, (iii) $\lim_{x \to +\infty} F(x) = 1$, (iv) $\lim_{n \to \infty} F(x + 1/n) = F(x)$.

Corollary 3.4. *If F is a distribution function then there exists a probability $\mathbf{Q}$ on $(\mathbb{R}, \mathscr{B})$ such that $\mathbf{Q}(-\infty, x] = F(x)$ for all $x \in \mathbb{R}$.*

Proof Let $\Omega = \{0,1\}^{\mathbb{N}}$. Let $\mathscr{F}$ be as in (3.4) or (3.5). Let $\mathbf{P}$ be the probability on $(\Omega, \mathscr{F})$ as in Theorem 3.1. Let $U : (\Omega, \mathscr{F}) \to (\mathscr{R}, \mathscr{B})$ be the random variable defined by $U(\omega) = \sum_{n=1}^{\infty} 2^{-n} \omega_n$. Let F^{-1} be defined by (3.8). Let $X = F^{-1} \circ U$. Then $\mathbf{Q} = \mathbf{P}_X$. □

Lemma 3.16. *Let X be one real random variable with law P_X and distribution function $F(x) = \mathbf{P}_X(-\infty, x]$, $x \in \mathbb{R}$. Then (i) $\mathbf{P}(X \in (a,b]) = F(b) - F(a)$, (ii) $\mathbf{P}(X \in (a,b)) = F(b-) - F(a)$, (iii) $\mathbf{P}(X \in [a,b]) = F(b) - F(a-)$, (iv) $\mathbf{P}(X = a) = F(a) - F(a-)$.*

Proof (i)

$$(a,b] = (-\infty, b] - (-\infty, a].$$

(ii)

$$(a,b) = \cup_{n \in \mathbb{N}} (a, b - 1/n]$$

(iii)

$$[a,b] = \cap_{n \in \mathbb{N}} (a + 1/n, b].$$

(iv)

$$\{a\} = [a,a].$$

□

EXERCISE 16. Carefully justify the formulae in the proof of Lemma 3.16.

EXERCISE 17. There is nothing kosher about choosing F so that it is right continuous. It is, merely, a convention. Another choice could be $F(x) = \mathbf{P}(-\infty, x)$, which (show this) results into a left continuous function. Yet another choice is to take this last function and modify it at each of points of discontinuity x and giving it the value $\frac{1}{2}(F(x-) + F(x+))$.

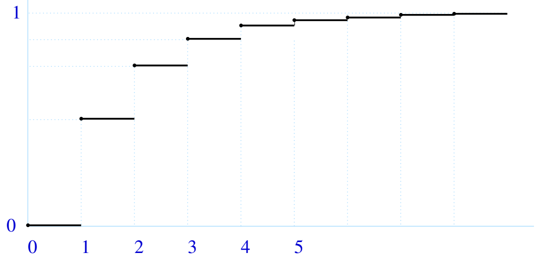
3.11 Types of distribution functions

In this section we discuss the various kinds of distribution functions on $\mathbb{R}$.

3.11.1 Discrete distribution functions

A discrete distribution function is the distribution function of a discrete random variable X with values in some countable subset S of $\mathbb{R}$. Assume that $p(s) = \mathbf{P}(X = s) > 0$ for all $s \in S$. Such a distribution function satisfies $F(s) - F(s-) > 0$ for all $s \in S$. Indeed, $F(s) - F(s-) = \mathbf{P}(X = s)$. Also, if (a,b) is an open interval containing no points of S, then F is constant on (a,b). Indeed, if $a < x < b$ then $F(x) - F(a) = \mathbf{P}(a < X \leq x) = \sum_{s \in S, a < s \leq x} \mathbf{P}(X = s) = 0$.

Example 3.4. Let X be a random variable such that $\mathbf{P}(X = n) = 2^{-n}$, $n \in \mathbb{N}$. Then its distribution function looks like

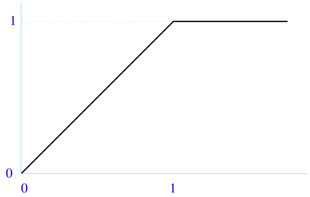


Example 3.5. Let X be a random variable such that for every rational number of the form m/n where m, n are integers with no common factors, we have $\mathbf{P}(X = m/n) = c2^{-(m+n)}$ where c is chosen so that $\mathbf{P}(X \in \mathbb{Q}) = 1$. Its distribution function is discrete because $\mathbb{Q}$ is countable. Unfortunately, I can't draw it. (There are no intervals (a, b) containing no rational points.)

3.11.2 Continuous distribution functions

A distribution function F is continuous if it is a continuous function, i.e. if $F(x) - F(x-) = 0$ for all $x \in \mathbb{R}$.

Example 3.6. Consider the random variable U with $\mathbf{P}(U \le u) = u$ for all $u \in [0, 1]$. Such a random variable exists (we constructed it). Its distribution function looks like



Excepting the points $0, 1$ we have that it is also differentiable with derivative $f(u) = 1$ if $0 < u < 1$ and 0 otherwise.. If we arbitrarily define $f(0) = f(1) = 0$, we also have $\int_{-\infty}^{u} f(t)dt = F(u)$ for all $u \in \mathbb{R}$. We like such distribution functions:

Absolutely continuous distribution functions

A distribution function F is called ABSOLUTELY CONTINUOUS if there exists a function f (called DENSITY of F) such that

$$F(x) = \int_{-\infty}^{x} f(t)dt, \quad x \in \mathbb{R}.$$

[e] The density is not uniquely defined. For instance, it can be changed on a finite set and such a change will not affect the integral above. Usually, one[f] imposes additional regularity conditions, such as continuity, resulting in uniqueness.

But not all continuous distribution functions are absolutely continuous:

Singularly continuous distribution functions

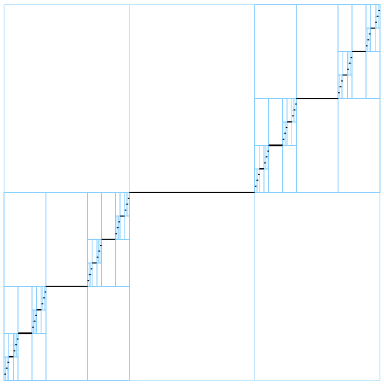
A distribution function F is called SINGULARLY CONTINUOUS if it is continuous but not absolutely continuous. We need to show that there are such functions.

Example 3.7. Consider the coin-flip space $(\Omega = \{0, 1\}^{\mathbb{N}}, \mathscr{F}, \mathbf{P})$ and let

$$V(\omega) := \sum_{n=1}^{\infty} \frac{2\omega_n}{3^n}.$$

[e] The integral in the display is a Lebesgue integral. For a definition, skip to §3.13.3. For the time being, you may think of it as the standard Riemann integral of Integral Calculus.

[f] unconsciously



EXERCISE 18. Show that the random variable V defined in Example 3.7 has a continuous but not absolutely continuous distribution function.

3.11.3 General distribution functions

Suppose that F, G are distribution functions. Then, for any $\lambda \in (0, 1)$, the function $\lambda F + (1 - \lambda G)$ is a distribution function. (Probabilistically, if X, Y are random variables with distribution functions F, G, respectively, then we can define a new random variable Z which equals X with probability λ or Y with probability $1 - \lambda$.) So, if F is discrete and G continuous then $\lambda F + (1 - \lambda G)$ is neither discrete nor continuous: it is mixed. The question is: Can we exhaust all distribution functions by taking mixtures of the three types mentioned above? The answer is yes:

Theorem 3.2. *Let F be a distribution function on $\mathbb{R}$. Then F can be* uniquely *written as*

$$F = \lambda_d F_d + \lambda_{ac} F_{ac} + \lambda_{sc} F_{sc}$$

where F_d, F_{ac}, F_{sc} are discrete, absolutely continuous, singularly continuous distribution functions, respectively, and where the coefficients are nonnegative such that $\lambda_d + \lambda_{ac} + \lambda_{sc} = 1$.

The last two terms of this decomposition are known as the continuous part of F. The first two terms are known as the singular part of F. We will not prove this theorem, but refer, e.g. to [2].

3.11.4 Differentiation: a word of caution

The subject of densities involves the concept of a derivative of functions that are not necessarily everywhere differentiable. Mimicking the definition of a density, we will say that a function G has density g if $G'(x) = g(x)$ for almost all x. The latter statement means that it holds true that $G'(x) = g(x)$ for all x in some set A whose complement is small in the sense that for all $\varepsilon > 0$ there exist intervals I_n, $n \in \mathbb{N}$ with lengths λ_n, $n \in \mathbb{N}$, such that $\sum_n \lambda_n < \varepsilon$ and $A^c \subset \cup_n I_n$. We take this latter statement as a definition:

Definition 3.2. (i) We say that a set $B \subset \mathbb{R}$ has MEASURE ZERO if all $\varepsilon > 0$ there exist intervals I_n, $n \in \mathbb{N}$ with lengths λ_n, $n \in \mathbb{N}$, such that $\sum_n \lambda_n < \varepsilon$ and $B \subset \cup_n I_n$.
(ii) A function $G : \mathbb{R} \to \mathbb{R}$ is said to be ALMOST EVERYWHERE DIFFERENTIABLE if $G'(x)$ exists for all $x \in A$ where A^c has measure zero.

We can verify the following:

- Any countable set has measure zero.

- A set $B \subset [0,1]$ has measure zero if and only if a uniform random variable U in $[0,1]$ satisfies $\mathbf{P}(U \in B) = 0$.

- An absolutely continuous distribution function is almost everywhere differentiable.

What kind of functions G are almost everywhere differentiable with a derivative g that can be used to recover the function? That is, for what kind of functions G can we apply the Fundamental Theorem of Calculus (FTC), i.e. the statement that $G(b) - G(a) = \int_a^b g(x)dx$? Calculus tells us that if G is piecewise continuously differentiable then this is true. However, we saw, by means of distribution functions, not only that the FTC holds more generally but it is desirable to understand it more generally because random variables with not everywhere differentiable distribution functions abound in theory and in practise. Here is a definition.

Definition 3.3. A function $G : \mathbb{R} \to \mathbb{R}$ is called ABSOLUTELY CONTINUOUS if for all $\varepsilon > 0$ there exists a $\delta > 0$ such that for any finite collection of disjoint intervals $[a_k, b_k]$ we have $\sum_k |G(b_k) - G(a_k)| < \varepsilon$ provided that the sum of the lengths of the intervals is less than δ.

EXERCISE 19. A differentiable function is absolutely continuous. An absolutely continuous function is continuous.

Theorem 3.3. *Any absolutely continuous function G is differentiable almost everywhere and if g is its derivative then the FTC holds.*

For a proof see [2]. Alternatively, one can use Probability Theory to prove all that, provided that one has understood the theory of Martingales, a subject of a later chapter. Indeed, much of this generalised differentiation theory achieves a beautiful interpretation and becomes comprehensible in probabilistic terms, via Martingale Theory.

EXERCISE 20. Show that a continuous and piecewise differentiable function G is almost everywhere differentiable.

3.12 Transformation rules and densities

Consider a random variable $X : (\Omega, \mathscr{F}) \to (\mathscr{R}, \mathscr{B})$. Suppose $\mathbf{P}$ is a probability on $(\Omega, \mathscr{F})$. We are interested in the distribution $\mathbf{P}_X$ of X. Suppose, for some reason, we don't like it and want to change it to something else. There are two ways to do this. First, we can change the probability $\mathbf{P}$ and replace it by some other probability $\mathbf{Q}$. Then $\mathbf{P}_X$ will be replaced by $\mathbf{Q}_X$. Second, we can take a function $H : (\mathscr{R}, \mathscr{B}) \to (\mathscr{R}, \mathscr{B})$ and replace X by $H \circ X$. Then $\mathbf{P}_X$ is replaced by $\mathbf{P}_{H \circ X}$. The two procedures are not, in general, equivalent.

Suppose, for instance, that X is a discrete random variable. Then any one-to-one function H will not change the probabilities of singletons $\{s\}$ such that $\mathbf{P}_X\{s\} > 0$, but, merely, will rename them: $\{s\}$ will be transformed to $\{H(s)\}$ and its probability will remain the same. Even if H is not one-to-one, there is not much that H can do to change the probabilities. Consider, for instance, a random variable X with values $1, 2, 3$ and probabilities p_1, p_2, p_3, respectively. Then the most a function H can do is either be one-to-one, in which case $H(1), H(2), H(3)$ will retain the old probabilities, or map two points, say $1, 2$, to a single point, with probability $p_1 + p_2$ and leave the the third intact. Thus, the types of changes in the distribution of a discrete X that can be achieved by taking a function of it are quite restricted. To really change its distribution *ad libitum*, we need to change the underlying probability $\mathbf{P}$.

For the case of absolutely continuous random variables, the story is different: a merely one-to-one function H can simultaneously change the values and the distribution in a quite general fashion.

Theorem 3.4. *Let X be an absolutely continuous random variable in $\mathbb{R}$ with density f. Let $\varphi : \mathbb{R} \to \mathbb{R}$ be strictly increasing differentiable function and let ψ be its inverse function. Then $\varphi(X)$ a random variable with absolutely continuous distribution function and density*

$$\psi' \cdot f \circ \psi \text{ on } \varphi(\mathbb{R}),$$

and 0 *elsewhere.*

Proof Since φ is strictly increasing, its inverse function exists and has domain $\varphi(\mathbb{R})$. Then, the distribution function of $\varphi(X)$ is, for any $t \in \varphi(\mathbb{R})$,

$$\mathbf{P}(\varphi(X) \le t) = \mathbf{P}(X \le \psi(t)) = \int_{-\infty}^{\psi(t)} f(x)dx.$$

By changing variable in the integral we have

$$\int_{-\infty}^{\psi(t)} f(x)dx = \int_{-\infty}^{t} f(\psi(s))\psi'(s)ds,$$

where we set $\psi'(s) = 0$ for $s \notin \varphi(\mathbb{R})$. From the definition of an absolutely continuous distribution function we see that, indeed, $\varphi(X)$ has absolutely continuous distribution function and its density is the function inside the last integral. □

EXERCISE 21. Let U be a uniform random variable in the interval $(0, 1)$. Find the density function of e^{e^U}.

Theorem 3.4 assumes that φ is strictly increasing. It is immediate to find out the formula for strictly decreasing φ. Generalising to more general functions is possible and relatively easy for random variables with values in $\mathbb{R}$ (the story in $\mathbb{R}^d$ is more complicated). For instance, we may assume that φ is piecewise differentiable. The problem becomes a problem in differential calculus and the general theorem is omitted. However, an example is due:

EXERCISE 22. Let X be a random variable with density $f(x) = c(1 + x^2)^{-1}$, $x \in \mathbb{R}$. Let $\varphi(x) = \cosh(x)$. Find the density (and hence show that it exists) of $\varphi(X)$.

3.13 Expectation

The expectation of one real random variable is, if it can be defined, an important numerical aspect of the random variable. It is justified, for instance, by the Theorem (Law) of Large Numbers which will be proved at a later chapter.

It is easy to define the expectation of a discrete random variable X with values in $\mathbb{R}$ and probability mass function $p(x)$. Let S be the set of x such that $p(x) > 0$. Then

$$\mathbf{E}X = \sum_{x \in S} x p(x),$$

provided that this sum can be defined. We know from Analysis (see [2]) that the sum of positive numbers can be defined irrespective of which order we sum the numbers up. However, not all the summands above are necessarily positive. So let us consider the positive and negative terms separately and try to define

$$\mathbf{E}X = \sum_{x \in S_+} x p(x) - \sum_{x \in S_-} (-x) p(x),$$

where $S_+ := \{x \in S : \ p(x) > 0\}$, where $S_- := \{x \in S : \ p(x) < 0\}$. Each of the two sums, separately, is a sum of positive terms, hence it is well-defined. The only "problem" is that such a sum can take value $+\infty$. If both sums are finite then $\mathbf{E}X$ is a finite number. If the first sum is $+\infty$ but the second finite then $\mathbf{E}X = +\infty$. Similarly, if the first is finite but the second infinite, then $\mathbf{E}X = -\infty$. The only case where we cannot talk (cannot define) $\mathbf{E}X$ is when both sums are infinite.

When X has absolutely continuous distribution function with density f, one can define $\mathbf{E}X$ similarly:

$$\mathbf{E}X = \int_0^\infty xf(x)dx - \int_{-\infty}^0 (-x)f(x)dx,$$

provided that not both integrals are infinity.

In the general case, we need to be more prudent. We will treat the general case, not only because there are random variables which are neither discrete or absolutely continuous but because we can seldom (very seldom indeed) rely on knowledge of the probability mass function or density function.

3.13.1 Definition of expectation

We are going to define the expectation of a random variable $X : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B})$ with respect to a probability $\mathbf{P}$ sitting on $(\Omega, \mathscr{F})$. We will use the notation $\mathbf{E}_\mathbf{P}X$ or, simply, $\mathbf{E}X$ is the probability $\mathbf{P}$ is understood from the context. After constructing this, we will see that it does not depend on the choice of probability space, but only on the law of X, namely, we shall show that if $X' : (\Omega', \mathscr{F}') \to (\mathbb{R}, \mathscr{B})$ is another random variable and $\mathbf{P}'$ a probability on $(\Omega', \mathscr{F}')$ such that $\mathbf{P}_X = \mathbf{P}'_{X'}$ then $\mathbf{E}_\mathbf{P}X = \mathbf{E}_{\mathbf{P}'}X'$.

▷ SIMPLE RANDOM VARIABLES

The simplest possible random variable is

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

Since $\mathbf{1}_A$ takes value 1 with probability $\mathbf{P}(A)$ or 0 with the complementary probability, it is only reasonable to define

$$\mathbf{E}\mathbf{1}_A = \mathbf{P}(A).$$

We "know" that the expectation should be linear. So we can extend this definition to linear combinations of indicators (simple random variable) and define

$$\mathbf{E}(a_1\mathbf{1}_{A_1} + \cdots + a_m\mathbf{1}_{A_m}) = a_1\mathbf{P}(A_1) + \cdots + a_m\mathbf{P}(A_m). \tag{3.9}$$

Is this a good definition? There is a slight issue here because a simple random variable can be written as linear combination of indicators in two different ways. For example

$$8\mathbf{1}_A - 3\mathbf{1}_B = 8\mathbf{1}_{A \setminus B} - 3\mathbf{1}_{B \setminus A} + 5\mathbf{1}_{AB}.$$

Using the left side we should define the expectation of it $8\mathbf{P}(A) - 3\mathbf{P}(B)$, while, using the right side, as $8\mathbf{P}(A \setminus B) - 3\mathbf{P}(B \setminus A) + 5\mathbf{P}(AB)$. So we must have

$$8\mathbf{P}(A) - 3\mathbf{P}(B) = 8\mathbf{P}(A \setminus B) - 3\mathbf{P}(B \setminus A) + 5\mathbf{P}(AB).$$

This is true because of additivity of $\mathbf{P}$. Indeed, write $\mathbf{P}(A) = \mathbf{P}(A \setminus B) + \mathbf{P}(AB)$, $\mathbf{P}(B) = \mathbf{P}(B \setminus A) + \mathbf{P}(AB)$, substitute in the left side, and you obtain the right side. This argument extends to the general case and so the definition (3.9) is good.

Lemma 3.17 (algebraic properties)**.** *Suppose ξ, η are simple random variables on the same probability space $(\Omega, \mathscr{F}, \mathbf{P})$. Then:*
(i) If $\mathbf{P}(A) = 0$ then $\mathbf{E}\xi\mathbf{1}_A = 0$.
(ii) $\mathbf{E}(c\xi) = c\mathbf{E}(\xi)$ for all $c \in \mathbb{R}$.
(iii) $\mathbf{E}(\xi + \eta) = \mathbf{E}\xi + \mathbf{E}\eta$.
(iv) If $\mathbf{P}(\xi \geq 0) = 1$ then $\mathbf{E}\xi \geq 0$. If $\mathbf{P}(\xi \leq \eta) = 1$ then $\mathbf{E}\xi \leq \mathbf{E}\eta$.

Proof (i) Suppose A is an event in $\mathscr{F}$ with $\mathbf{P}(A) = 0$. Then $\mathbf{E}\xi\mathbf{1}_A = 0$ because: If $\xi = \sum_k a_k \mathbf{1}_{A_k}$ then $\xi\mathbf{1}_A = \sum_k a_k \mathbf{1}_{A_k A}$ and so $\mathbf{E}\xi\mathbf{1}_A = \sum_k a_k \mathbf{P}(AA_k)$. But $\mathbf{P}(AA_k) \leq \mathbf{P}(A) = 0$, for all k.
(ii), (iii), (iv) (Almost) immediate from the definition (3.9). □

We now concentrate to nonnegative simple random variables. We allow such a random variable to take value $+\infty$. In other words, ξ is a nonnegative simple random variable if $\xi = \sum_{i=1}^m a_i \mathbf{1}_{A_i}$, with $a_i \in \mathbb{R} \cup \{+\infty\}$ for all i. The expectation of such a random variable is defined as before: $\mathbf{E}\xi = \sum_{i=1}^m a_i \mathbf{P}(A_i)$ and the possibility that one of the values may be equal to $+\infty$ worries us not, for all numbers are nonnegative, so the worst that can happen is that $\mathbf{E}\xi$ may be $+\infty$.

Lemma 3.18 (analytical properties)**.** *Suppose $\xi, \xi_1, \xi_2, \ldots,\ \eta_1, \eta_2, \ldots$ are nonnegative simple random variables, and $A \in \mathscr{F}$. Then:*
(i) If ξ_n is an increasing sequence with limit $\mathbf{1}_A$ then $\mathbf{E}\xi_n$ has limit $\mathbf{P}(A)$.
(ii) If ξ_n is an increasing sequence with limit ξ then $\mathbf{E}\xi_n$ has limit $\mathbf{E}\xi$.
(iii) If both ξ_n, η_n are increasing sequences such that $\lim_n \xi_n(\omega) = \lim_n \eta_n(\omega)$ for all $\omega \in \Omega$, then $\mathbf{E}\xi_n$, $\mathbf{E}\eta_n$ converge to the same limit.

Proof (i) Let $\varepsilon > 0$. If $\omega \in A$, then $\xi_n(\omega)$ converges to 1 and so, for all large n, $\xi_n(\omega) > 1 - \varepsilon$. In other words, the sequence of events $\{\xi_n > 1 - \varepsilon\}$ is increasing with union A. Hence $\mathbf{P}(\xi_n > 1 - \varepsilon)$ increases and has limit $\mathbf{P}(A)$. But

$$\mathbf{1}_A \geq \xi_n \geq (1 - \varepsilon)\mathbf{1}(\xi_n > 1 - \varepsilon).$$

Therefore, by Lemma 3.17(iv),

$$\mathbf{P}(A) \geq \mathbf{E}\xi_n \geq (1 - \varepsilon)\mathbf{P}(\xi_n > 1 - \varepsilon).$$

Since the last term converges to $\mathbf{P}(A)$ and since ε is arbitrary, we have that $\mathbf{E}\xi_n$ converges to $\mathbf{P}(A)$.
(ii) Let $a_1, a_2, \ldots, a_m$ be the distinct nonzero values of ξ. Then $\xi = \sum_{i=1}^m a_i \mathbf{1}(\xi = a_i)$. Since $\xi_n \uparrow \xi$, we have $\xi_n \mathbf{1}(\xi = a_i) \uparrow \xi\mathbf{1}(\xi = a_i) = a_1 \mathbf{1}(\xi = a_i)$, and so, $a_i^{-1}\xi_n \mathbf{1}(\xi = a_i) \uparrow \mathbf{1}(\xi = a_i)$. Applying (i), we have $\mathbf{E}[a_i^{-1}\xi_n \mathbf{1}(\xi = a_i)] \uparrow P(\xi = a_i)$, and so, by Lemma 3.17(ii), $\mathbf{E}[\xi_n \mathbf{1}(\xi = a_i)] \uparrow a_i P(\xi = a_i)$, for all $i = 1, \ldots, m$. Summing over i, we have

$$\sum_{i=1}^m \mathbf{E}[\xi_n \mathbf{1}(\xi = a_i)] \uparrow \sum_{i=1}^m a_i P(\xi = a_i).$$

By Lemma 3.17(iii), the left side equals $\mathbf{E}\xi_n$, while, by the definition (3.9), the right side equals $\mathbf{E}\xi$.
(iii) Since both sequences are increasing and have the same limit we have

$$\lim_{m\to\infty} (\xi_n \wedge \eta_m) = \xi_n, \quad \lim_{n\to\infty} (\xi_n \wedge \eta_m) = \eta_m.$$

By (ii), we have

$$\lim_{m\to\infty} \mathbf{E}(\xi_n \wedge \eta_m) = \mathbf{E}\xi_n, \quad \lim_{n\to\infty} \mathbf{E}(\xi_n \wedge \eta_m) = \mathbf{E}\eta_m.$$

Since the numerical sequence $\mathbf{E}(\xi_n \wedge \eta_m)$ is increasing in both arguments, we have that the limits over n and over m can be interchanged and so $\lim_{n\to\infty} \mathbf{E}\xi_n = \lim_{m\to\infty} \mathbf{E}\eta_m$. □

▷ NONNEGATIVE RANDOM VARIABLES

Suppose that X is a nonnegative random variable that may, possibly, take value $+\infty$. We define

$$\mathbf{E}X := \sup\{\mathbf{E}\xi : \ \xi \text{ is simple nonnegative, and } \xi \leq X\}.$$

We immediately have that there exists a sequence η_n of simple random variables such that $\eta_n \leq X$ for all n and $\mathbf{E}\eta_n \to \mathbf{E}X$.

Lemma 3.19. *For ANY sequence ξ_n of simple nonnegative random variables such that $\xi_n \uparrow X$, we have $\mathbf{E}\xi_n \uparrow \mathbf{E}X$.*

Proof As we just pointed out, there is one sequence η_n of nonnegative simple RVs such that $\eta_n \leq X$ and $\mathbf{E}\eta_n \uparrow \mathbf{E}X$. Let ξ_n be a sequence of simple random nonnegative variables such that $\xi_n \uparrow X$. Define $\zeta_n = \max(\eta_1, \ldots, \eta_n, \xi_n)$. Clearly, ζ_n is simple, and $\zeta_n \leq X$. Also, ζ_n is increasing with limit X. Since both ξ_n and ζ_n increase to the same limit, by Lemma 3.18(iii), $\mathbf{E}\xi_n$ and $\mathbf{E}\zeta_n$ have the same limit. But $\eta_n \leq \zeta_n \leq X$. Hence $\mathbf{E}\eta_n \leq \mathbf{E}\zeta_n \leq \mathbf{E}X$. Since $\mathbf{E}\eta_n$ has limit $\mathbf{E}X$, we conclude that $\mathbf{E}\zeta_n$ has limit $\mathbf{E}X$. Therefore $\mathbf{E}\xi_n$ has limit $\mathbf{E}X$ also. □

In the course of the proof we said that we can find a sequence of simple functions that increases to X. This is done through, for example, the functions τ_n defined by (3.1).

$$\tau_n(x) := 2^{-n}\lceil 2^n x\rceil \wedge n.$$

Since

$$\lceil x\rceil = \sum_{k=1}^{\infty} \mathbf{1}(k-1 \leq x < k), \quad x \geq 0$$

we can write

$$\tau_n(x) := \sum_{k=1}^{n2^n} 2^{-n}k\mathbf{1}(k-1 \leq 2^n x < k), \quad x \geq 0.$$

Each τ_n is left-continuous and $\tau_n(x) \to x$ as $n \to \infty$ for each x. Moreover, $\tau_n(x) \leq \tau_{n+1}(x)$ for all x. Therefore, for any nonnegative RV X, we have that $\tau_n(X)$ is an increasing sequence of simple nonnegative RVs with $\tau_n(X) \uparrow X$ and, by Lemma 3.19,

$$\mathbf{E}X = \lim_{n\to\infty} \sum_{k=1}^{n2^n} \frac{k}{2^n}\mathbf{P}\Big(\frac{k-1}{2^n} \leq X < \frac{k}{2^n}\Big).$$

Extending Lemma 3.19, we obtain

Theorem 3.5 (monotone convergence theorem)**.** *For ANY sequence X_n of nonnegative random variables such that $X_n \uparrow X$, we have $\mathbf{E}X_n \uparrow \mathbf{E}X$.*

Proof By Lemma 3.19, $\mathbf{E}X = \lim_{m\to\infty}\mathbf{E}\tau_m(X)$. By the left-continuity of τ_m, we have $\tau_m(X_n) \uparrow \tau_m(X)$, as $n \to \infty$, and by Lemma 3.19, the same is true after taking expectations. So:

$$\mathbf{E}X = \lim_{m\to\infty}\lim_{n\to\infty}\mathbf{E}\tau_m(X_n).$$

Because the numerical sequence $\mathbf{E}\tau_m(X_n)$ is increasing in both indices m and n, we can interchange the limits:

$$\mathbf{E}X = \lim_{n\to\infty}\lim_{m\to\infty}\mathbf{E}\tau_m(X_n) = \lim_{n\to\infty}\mathbf{E}X_n,$$

where the last equality follows, again, from Lemma 3.19. □

Lemma 3.20. *Let X, Y be nonnegative random variables on $(\Omega, \mathscr{F}, \mathbf{P})$. Then:*
(i) $\mathbf{E}X = 0$ *if and only if* $\mathbf{P}(X > 0) = 0$.
(ii) $\mathbf{E}(cX) = c\mathbf{E}X$ *for all* $c \geq 0$.
(iii) $\mathbf{E}(X + Y) = \mathbf{E}X + \mathbf{E}Y$.
(iv) If $\mathbf{P}(X \leq Y) = 1$ *then* $\mathbf{E}X \leq \mathbf{E}Y$.

Proof We only prove half of (i). The rest are easy. We have $\{X > 0\}$ is the limit of $\{X > 1/n\}$, as $n \to \infty$. Hence $P(X > 0) = \lim_{n\to\infty} P(X > 1/n)$. If $P(X > 0) > 0$ then $P(X > 1/n) > 0$ for some n and, since $X \geq (1/n)\mathbf{1}(X > 1/n)$ we have $\mathbf{E}X \geq (1/n)\mathbf{P}(X > 1/n)$, i.e. $\mathbf{E}X > 0$. □

Lemma 3.21 (Fatou's lemma)**.** *For any sequence X_n of nonnegative random variables,* $\mathbf{E}\,\underline{\lim}\,X_n \leq \underline{\lim}\,\mathbf{E}X_n$.

Proof We have

$$\inf_{k\geq n} X_k \leq X_\ell, \quad \ell \geq n.$$

So, by (iv) of Lemma 3.20,

$$\mathbf{E}\inf_{k\geq n} X_k \leq \mathbf{E}X_\ell, \quad \ell \geq n,$$

and so

$$\mathbf{E}\inf_{k\geq n} X_k \leq \inf_{\ell\geq n} \mathbf{E}X_\ell.$$

Both sides are increasing numerical sequences in n. The right side increases to $\underline{\lim}\,\mathbf{E}X_\ell$. The left side increases, by Theorem 3.5, to $\mathbf{E}\,\underline{\lim}\,X_n$. □

▷ GENERAL RANDOM VARIABLES

Suppose now X is a random variable with values in $\mathbb{R}$. We say that X has an expectation if not both $X^+ = \max(X, 0)$ and $X^- := -(-X)^+$ have infinite expectations. In this case, we define

$$\mathbf{E}X = \mathbf{E}X^+ - \mathbf{E}X^-.$$

The definition is, of course, well-motivated: for any number x, we have $x = x^+ - x^-$.

EXERCISE 23. If x is a real number, we let, as usual, $x^+ := \max(x, 0)$, $x^- := -(-x)^+$. Show that $x^- = -\min(x, 0)$ and derive the identities

$$x = x^+ - x^-, \quad |x| = x^+ + x^-, \quad \min(|x|, |y|) = \tfrac{1}{2}(|x| + |y| - |x - y|).$$

Show also that if $a < b$ then $(x \vee a) \wedge b = (x \wedge b) \vee a$.

We say that X is integrable (with respect to $\mathbf{P}$) if both $\mathbf{E}X^+$ and $\mathbf{E}X^-$ are finite or, equivalently, if $\mathbf{E}|X| < \infty$. (The latter follows from the identity $|x| = x^+ + x^-$.)

Lemma 3.22. *Let X, Y be integrable random variables on $(\Omega, \mathscr{F}, \mathbf{P})$. Then:*
(i) $\mathbf{E}(cX) = c\mathbf{E}X$ *for all* $c \in \mathbb{R}$.
(iii) $\mathbf{E}(X + Y) = \mathbf{E}X + \mathbf{E}Y$.
(iv) If $\mathbf{P}(X \leq Y) = 1$ *then* $\mathbf{E}X \leq \mathbf{E}Y$.

If $A \in \mathscr{F}$ we can define the expectation of X on A by:

$$\mathbf{E}(X; A) := \mathbf{E}(X\mathbf{1}_A).$$

If $\mathbf{P}(A) > 0$ we can define the expectation of X given A by:

$$\mathbf{E}(X|A) := \frac{\mathbf{E}(X;A)}{\mathbf{P}(A)}.$$

We remark that $\mathbf{E}(X|A)$ is expectation with respect to the restriction $\mathbf{P}_A$ of $\mathbf{P}$ on A, i.e. with respect to the probability

$$\mathbf{P}_A : \mathscr{F} \to \mathbb{R}; \qquad \mathbf{P}_A(B) := \mathbf{P}(AB), \quad B \in \mathscr{F}.$$

In other words,

$$\mathbf{E}(X|A) = \mathbf{E}_{\mathbf{P}_A} X.$$

EXERCISE 24. Show that $|\mathbf{E}X| \le \mathbf{E}|X|$ (whenever $\mathbf{E}X$ is defined).

Theorem 3.6 (Dominated Convergence Theorem)**.** *Let X_n be a sequence of random variables such that $X(\omega) = \lim_{n\to\infty} X_n(\omega)$ exists and such that $|X_n(\omega)| \le Y(\omega)$ for all n and ω, and $\mathbf{E}|Y| < \infty$. Then $\mathbf{E}|X_n - X|$ converges to zero.*

Proof Apply Fatou's lemma 3.21 to $2Y - |X_n - X|$. □

3.13.2 Substitution rule

Lemma 3.23. *Consider the measurable functions*

$$(\Omega, \mathscr{F}) \xrightarrow{Z} (S, \mathscr{S}) \xrightarrow{H} (\mathbb{R}, \mathscr{B}).$$

Let $\mathbf{P}$ be a probability on $(\Omega, \mathscr{F})$. Let $\mathbf{P}_Z$ be the law of Z. Then

$$\mathbf{E}_{\mathbf{P}} H \circ Z = \mathbf{E}_{\mathbf{P}_Z} H,$$

whenever wither side exists. (Here, $H \circ Z$ is one real random variable on the probability space $(\Omega, \mathscr{F}, \mathbf{P})$ and H is one real random variable on the probability space $(S, \mathscr{S}, \mathbf{P}_Z)$.)

Proof Suppose H is an indicator random variable, i.e. $H = \mathbf{1}_B$ for some $B \in \mathscr{S}$. Then $\mathbf{E}_{\mathbf{P}_Z} \mathbf{1}_A = \mathbf{P}_Z(A)$ by the definition of the expectation of a simple random variable. On the other hand, $H \circ Z = \mathbf{1}_A(Z)$ is an indicator random variable on $(\Omega, \mathscr{F})$: it is the indicator of the set $\{\omega \in \Omega : Z(\omega) \in A\}$. Hence, again by the by the definition of the expectation of a simple random variable, $\mathbf{E}_{\mathbf{P}} H \circ Z = \mathbf{P}(Z \in A)$. But $\mathbf{P}_Z(A) = \mathbf{P}(Z \in A)$ by the definition of the law of the random variable Z (see section 3.6). Suppose next that H is a simple random variable. Use the above and linearity of expectation to get the result. Suppose that H is a nonnegative random variable. Use Lemma 3.5. Finally, suppose that H has no sign restriction, and use the definition of expectation. □

Corollary 3.5. *If X is a real random variable on $(\Omega, \mathscr{F}, \mathbf{P})$ with expectation $\mathbf{E}_{\mathbf{P}} X$ and law $\mathbf{P}_X$ then*

$$\mathbf{E}_{\mathbf{P}} X = \mathbf{E}_{\mathbf{P}_X} \iota$$

where $\iota : \mathbb{R} \to \mathbb{R}$ is the identity function: $\iota(x) \equiv x$.

Therefore, the expectation of a random variable depends only on its law.

Suppose that X is a discrete random variable with values in a finite set $S \subset \mathbb{R}$ and probability mass function $p(x), x \in S$. Since $X = \sum_{x \in S} x \mathbf{1}(X = x)$ is a simple random variable, we immediately have that $\mathbf{E}X = \sum_{x \in S} x \mathbf{P}(X = x) = \sum_{x \in S} x p(x)$, as needed. The same formula holds for a discrete random variable with values in a countable set S: Simply enumerate the elements of S and use monotone convergence theorem.

Let us now consider an absolutely continuous random variable X with density f. We would like to show that $\mathbf{E}X$ is compatible with the definition given at the beginning of the section, namely that it equals $\int_{\mathbb{R}} x f(x) dx$. To do this, we need to revisit the notion of an integral:

3.13.3 Expectation and densities

Let $h : (\mathbb{R}, \mathscr{B}) \to (\mathbb{R}, \mathscr{B})$ be measurable and suppose $h \geq 0$. Consider the random variable U on the coin-flip space $(\Omega = \{0,1\}^{\mathbb{N}}, \mathscr{F}, \mathbf{P})$, defined by (3.6). We define $\int_{\mathbb{R}} h(x)dx$ by

$$\int_{\mathbb{R}} h(x)dx := \sum_{n \in \mathbb{Z}} \mathbf{E} h(U+n).$$

This is called the Lebesgue integral of h. We also define $\int_a^b h(x)dx$ by

$$\int_a^b h(x)dx := \int_{\mathbb{R}} h(x)\mathbf{1}(a \leq x \leq b)dx,$$

and, more generally, for any $B \in \mathscr{B}$,

$$\int_B h(x)dx := \int_{\mathbb{R}} h(x)\mathbf{1}(x \in B)dx.$$

If h has no restriction on sign, we define $\int_a^b h(x)dx$ as $\int_a^b h^+(x)dx - \int_a^b h^-(x)dx$ provided that not both terms are infinite. This integral behaves like the ordinary Riemann integral for "nice" functions (e.g. for continuous or piecewise continuous functions).

More generally, if $F : \mathbb{R} \to \mathbb{R}$ is an increasing function, we define

$$\int_{\mathbb{R}} h(x)F(dx) = \int_{\mathbb{R}} h(F^{-1}(x))dx,$$

where $F^{-1}(x) := \inf\{t \in \mathbb{R} : \ F(t) > x\}$. This is the Lebesgue-Stieltjes integral. It is also denoted by $\int_{\mathbb{R}} h(x)dF(x)$ or simply as $\int_{\mathbb{R}} hdF$ when the variable of integration needs no mentioning.

> **The following theorem connects what one learns in basic Calculus with what we just defined. You absolutely need this theorem in order to make the connection between the standard Calculus tricks and recipes[g] and what we are talking about here.**

Theorem 3.7. *(i) If h is Riemann integrable on $[a,b]$ then the Lebesgue integral $\int_a^b h(x)dx$ exists and coincides with the Riemann integral.*
(ii) If h is bounded and measurable then it is Riemann integrable on $[a,b]$ if the set of discontinuities of h have measure zero, in the sense of Definition 3.2.

Example 3.8. For example, if h is piecewise continuous, then its discontinuities have measure zero and so it is Riemann integrable.

Example 3.9. As a counterexample, consider the function $h(x) = \mathbf{1}(x \in \mathbb{Q})$, $0 \leq x \leq 1$. This function is discontinuous on every point of $[0,1]$. The interval $[0,1]$ does not have measure zero. Hence it is not Riemann integrable. However, h is measurable and nonnegative, and so it does have a Lebesgue integral: this integral equals $\mathbf{E}h(U) = \mathbf{P}(U \in \mathbb{Q}) = 0$.

Theorem 3.7 enables us to use all rules we learn in basic Integration Calculus to the case of Lebesgue integration when the two integrals coincide.

Suppose now that X is a random variable with absolutely continuous distribution function F and density f. Consider the function $\mathbf{Q} : \mathscr{B} \to \mathbb{R}$ defined by

$$\mathbf{Q}(B) := \int_B f(x)dx.$$

Since $\mathbf{Q}(B) = \mathbf{P}_X(B)$ for $B = (-\infty, x]$, we have that $\mathbf{Q} = \mathbf{P}_X$. More generally,

[g] I mean things like $\int_0^x t^2 e^t dt = x^2 e^x - 2\int_0^x te^t dt = (x^2 - 2x + 2)e^x - 2$.

Lemma 3.24. *Let $\mathbf{P}_X$ be the law of a random variable X with density f. Then, for any measurable $g : (\mathbb{R}, \mathscr{B}) \to (\mathbb{R}, \mathscr{B})$,*

$$\int_{\mathbb{R}} g(x)f(x)dx = \mathbf{E}_{\mathbf{P}_X} g$$

provided that wither side exists.

Proof If $g = \mathbf{1}_B$, then this is what was discussed before the Lemma. For g simple, we use linearity. For general g we approximate. □

Corollary 3.6. *Let $\mathbf{P}_X$ be the law of an integrable random variable X with density f.*
(i) If ι is the identity function on $\mathbb{R}$ then

$$\int_{\mathbb{R}} xf(x)dx = E_{\mathbf{P}_X} \iota.$$

(ii)

$$\mathbf{E}X = \int_{\mathbb{R}} xf(x)dx.$$

Proof (i) follows from Lemma 3.24 with $g = \iota$. (ii) follows from corollary 3.5 □

EXERCISE 25. Suppose that Z is a real random variable with absolutely continuous distribution function and density f_Z. Let $H : (\mathbb{R}, \mathscr{B}) \to (\mathbb{R}, \mathscr{B})$ be a measurable function. Suppose that the random variable $X = H(Z)$ has density f_X. Show that the expectation of X (if it exists) can be computed in two ways:

$$\mathbf{E}X = \int_{\mathbb{R}} xf_X(x)dx = \int_{\mathbb{R}} H(z)f_Z(z)dz.$$

EXERCISE 26. Let $B \in \mathscr{B}$ and let $\lambda(B) := \int_{\mathbb{R}} \mathbf{1}_B(x)dx$ (Lebesgue integral). Show that $\lambda : \mathscr{B} \to \mathbb{R}$ satisfies $\lambda(\cup_n A_n) = \sum_n \lambda(A_n)$ whenever the A_n are mutually disjoint elements of $\mathscr{B}$ and that $\lambda(A + t) = \lambda(A)$ for all $A \in \mathscr{B}$, $t \in \mathbb{R}$, where $A + t := \{a + t : \ a \in A\}$. The function λ is called LENGTH.

EXERCISE 27. Consider a compass with a laser pointer attached at both ends of the needle. Suppose there is an infinite screen at some distance from the compass. Give it a spin and see mark X the location of the light with respect to a fixed point O on the screen (positive if it is to the right of O; negative if it is to the left). Show that $\mathbf{E}X$ is not defined. (You first must translate this problem in Mathematics.)

EXERCISE 28. Consider the function $F(x) := \sum_{k\in\mathbb{N}} \mathbf{1}(x \le k)$, $x \ge 0$, $F(x) = 0$, $x < 0$. Show that

$$\int_{\mathbb{R}} h(x)F(dx) = \sum_{k=1}^{\infty} h(k),$$

whenever the sum on the right makes sense.

3.14 Inequalities

Mathematics needs inequalities probably more than equalities. Probability, in particular, which, in some sense, contains a lot of approximation ideas needs inequalities. This section discusses some basic ones.

3.14.1 Markov, Chebyshev, Chernoff

Lemma 3.25 (Markov inequality). *If X is a nonnegative random variable then*

$$\mathbf{P}(X \geq t) \leq \frac{\mathbf{E}X}{t}, \quad t > 0.$$

Proof We have

$$t\mathbf{1}(X \geq t) \leq X$$

and $\mathbf{E}$ is preserved by $\leq$. □

Definition 3.4. The variance of a real random variable X with $\mathbf{E}X^2 < \infty$ is defined by

$$\operatorname{var} X := \mathbf{E}(X - \mathbf{E}X)^2.$$

Lemma 3.26 (Chebyshev inequality). *If X is a real random variable with $\mathbf{E}X^2 < \infty$ then*

$$\mathbf{P}(|X - \mathbf{E}X| \geq t) \leq \frac{\operatorname{var} X}{t}$$

Proof Apply the Markov inequality to $|X - \mathbf{E}X|$. □

Lemma 3.27 (Chernoff inequality). *If X is a real random variable then*

$$\mathbf{P}(X \geq t) \leq \frac{\mathbf{E}g(X)}{g(t)}$$

where g is a positive increasing function.

Proof Since g is increasing,

$$\{X \geq t\} \subset \{g(X) \geq g(t)\}$$

Now apply the Markov inequality to $g(X)$. □

EXERCISE 29. Let X be a discrete random variable with $\mathbf{P}(X = k) = \binom{n}{k}2^{-k}$, $k = 0, 1, \ldots, n$. Estimate $\mathbf{P}(X > na)$ for $a > 0.5$ using the above inequalities.

3.14.2 Jensen

A function $\varphi : \mathbb{R} \to \mathbb{R}$ is convex if

$$\varphi(pa + (1-p)b) \leq p\varphi(a) + (1-p)\varphi(b)$$

for all $a, b \in \mathbb{R}$ and all $0 \leq p \leq 1$. Notice that if ξ is a random variable with $\mathbf{P}(\xi = a) = p$, $\mathbf{P}(\xi = b) = 1 - p$, this definition can be written as

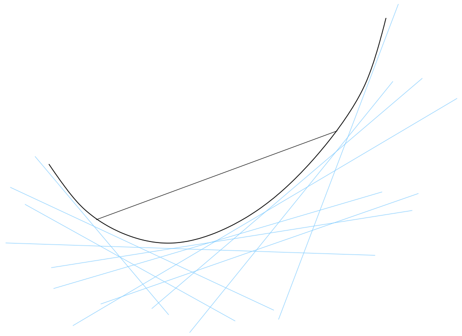
$$\varphi(\mathbf{E}\xi) \leq \mathbf{E}\varphi(\xi).$$

Jensen's inequality generalises this observation:

Lemma 3.28. *Let X be a real integrable random variable and φ a convex function. Then*

$$\varphi(\mathbf{E}X) \leq \mathbf{E}\varphi(X).$$

Geometrically, a convex function is such that the graph of the function on the interval $[a, b]$ lies below the straight segment with endpoints $(a, \varphi(a))$, $(b, \varphi(b))$.



If we take a straight line with slope equal to the slope of this segment and move it down, at some point it will be entirely below the whole graph of φ. Moreover, if we consider the set of all straight lines that are below the graph of φ and take their envelope (the maximum) then this equals φ. Namely,

$$\varphi(x) = \max_{\ell \in L(\varphi)} \ell(x), \quad x \in \mathbb{R}, \tag{3.10}$$

where $L(\varphi)$ is the collection of all functions $\ell(x) \equiv \alpha x + \beta$ such that $\ell \le \varphi$. To show this rigorously, observe that

$$a < b < c \Rightarrow \frac{\varphi(b) - \varphi(a)}{b - a} \le \frac{\varphi(c) - \varphi(b)}{c - b},$$

by the definition of convexity. Let $\varepsilon > 0$, $a = b - \varepsilon$, $c = b + \varepsilon$ and so

$$D_-\varphi(b) := \lim_{\varepsilon \downarrow 0} \frac{\varphi(b) - \varphi(b - \varepsilon)}{\varepsilon} \le \lim_{\varepsilon \downarrow 0} \frac{\varphi(b + \varepsilon) - \varphi(b)}{\varepsilon} := D_+\varphi(b),$$

where the limits exist by monotonicity. Let α be between $D_-\varphi(b)$ and $D_+\varphi(b)$. Then, if $x > b+\varepsilon$, we have

$$\frac{\varphi(b+\varepsilon) - \varphi(b)}{\varepsilon} \le \frac{\varphi(x) - \varphi(b)}{x - b},$$

and so

$$\alpha \le D_+\varphi(b) \le \frac{\varphi(x) - \varphi(b)}{x - b}, \quad x > b,$$

or $\alpha(x - b) + \varphi(b) \le \varphi(x)$, for all $x \ge b$. Arguing similarly for all $x \le b$, we have

$$\alpha(x - b) + \varphi(b) \le \varphi(x), \quad x \in \mathbb{R}.$$

Thus, the function $\ell(x) := \alpha(x - b) + \varphi(b) \le \varphi(x)$ belongs to $L(\varphi)$ and also, $\ell(b) = \varphi(b)$. This proves (3.10).

Proof of Lemma 3.28:

$$\varphi(\mathbf{E}X) = \max_{\ell \in L(\varphi)} \ell(\mathbf{E}X) = \max_{\ell \in L(\varphi)} \mathbf{E}\ell(X) \le \mathbf{E} \max_{\ell \in L(\varphi)} \ell(X) = \mathbf{E}\varphi(X).$$

The first equality is due to (3.10). The second uses linearity of expectation. The third uses monotonicity of expectation. The last reuses (3.10). □

EXERCISE 30. Let $a_1, \ldots, a_n$ be positive real numbers. Define their arithmetic, geometric and harmonic mean by

$$A_n = \frac{a_1 + \cdots + a_n}{n}, \quad G_n = (a_1 \cdots a_n)^{1/n}, \quad H_n = \frac{n}{a_1^{-1} + \cdots + a_n^{-1}},$$

respectively, and show that $A_n \ge G_n \ge H_n$.

3.15 Moments

Definition 3.5. When $r > 0$, the r-MOMENT of a nonnegative RV X is defined as the quantity $\mathbf{E}X^r$. The r-norm of a real RV X is defined as $||X||_r := (\mathbf{E}|X|^r)^{1/r}$.

Lemma 3.29. *The r-norm of X is increasing in r.*

Proof Let $r < s$ and $\varphi(x) = x^{s/r}$, $x > 0$. Notice that φ is convex. (This follows from the fact that its second derivative is positive.) Now apply the Jensen inequality. □

Corollary 3.7. *If $\mathbf{E}|X|^p < \infty$ for some $p > 0$ then $\mathbf{E}|X|^r < \infty$ for all $0 < r < p$.*

3.16 Hölder, Minkowski and Cauchy-Bunyakowskii-Schwarz

Definition 3.6. If X, Y are real random variables on the same $(\Omega, \mathscr{F}, \mathbf{P})$, the quantity $\mathbf{E}(XY)$ (whenever it is defined) is called CORRELATION between X and Y. The quantity $\text{cov}(X, Y) := \mathbf{E}((X - \mathbf{E}X)(Y - \mathbf{E}Y))$ is called COVARIANCE between X and Y.

Lemma 3.30 (Hölder inequality). *Let X, Y be real random variables. Then*

$$|\mathbf{E}(XY)| \le ||X||_p ||Y||_q,$$

for any $p, q > 0$, $p^{-1} + q^{-1} = 1$, as long as all terms involved exist and are finite.

Proof Let $(\Omega, \mathscr{F}, \mathbf{P})$ be a probability space on which both X, Y are defined. Without loss of generality assume that they are both nonnegative. Let $q > 1$ and assume that $\mathbf{E}(Y^q) < \infty$. Consider the probability

$$\mathbf{P}_q(A) := \frac{\mathbf{E}(Y^q \mathbf{1}_A)}{\mathbf{E}(Y^q)}, \quad A \in \mathscr{F}.$$

Let $\mathbf{E}_q$ denote expectation with respect to $\mathbf{P}_q$. Therefore, for any nonnegative random variable $W : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B})$,

$$\mathbf{E}_q W = \frac{\mathbf{E}(Y^q W)}{\mathbf{E}(Y^q)}. \tag{3.11}$$

Letting, in (3.11), $W = XY^{1-q}$, we obtain

$$\mathbf{E}(XY) = \mathbf{E}(Y^q)\ \mathbf{E}_q(XY^{1-q}).$$

Let p be defined from $p^{-1} + q^{-1} = 1$. Necessarily, $p > 1$. From Lemma 3.29 we have

$$\mathbf{E}_q Z \le (\mathbf{E}_q Z^p)^{1/p},$$

for any nonnegative random variable $Z : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B})$ with $\mathbf{E}Z^p < \infty$. Therefore,

$$\begin{aligned}
\mathbf{E}(XY) &\le \mathbf{E}(Y^q)\ (\mathbf{E}_q((XY^{1-q})^p))^{1/p} \\
&= \mathbf{E}(Y^q)\ \left(\frac{\mathbf{E}(X^p Y^q Y^{(1-q)p})}{\mathbf{E}(Y^q)}\right)^{1/p} \\
&= (\mathbf{E}(Y^q))^{1-1/p}\ (\mathbf{E}(X^p Y^{q+(1-q)p}))^{1/p}.
\end{aligned}$$

Since $1 - 1/p = 1/q$ and $q + (1 - q)p = 0$, the result follows. □

Corollary 3.8. (Cauchy-Bunyakowskii-Schwarz) *Let X, Y be real random variables. Then*

$$|\mathbf{E}(XY)| \leq ||X||_2||Y||_2,$$

as long as all terms involved exist and are finite.

Proof Notice that $\frac{1}{2} + \frac{1}{2} = 1$ and apply Hölder. □

Corollary 3.9. *Let X, Y be real random variables. Let*

$$\rho(X, Y) := \operatorname{cov}(X, Y)/\sqrt{\operatorname{var}(X)}\sqrt{\operatorname{var}(Y)},$$

whenever the terms exist. Then

$$-1 \leq \rho(X, Y) \leq 1.$$

Lemma 3.31 (Minkowski inequality)**.** *Let X, Y be real random variables. Then*

$$||X + Y||_p \leq ||X||_p + ||Y||_p,$$

for any $p > 1$, as long as all terms involved exist and are finite.

Proof Use the Hölder inequality as follows:

$$\begin{aligned}\mathbf{E}(|X+Y|^p) &= \mathbf{E}(|X|\ |X+Y|^{p-1}) + \mathbf{E}(|Y|\ |X+Y|^{p-1}) \\ &\leq [\mathbf{E}(|X|^p)]^{1/p}\ [\mathbf{E}(|X+Y|^{(p-1)q})]^{1/q} + [\mathbf{E}(|Y|^p)]^{1/p}\ [\mathbf{E}(|X+Y|^{(p-1)q})]^{1/q} \\ &= (||X||_p + ||Y||_p)\ [\mathbf{E}(|X+Y|^p)]^{1/q},\end{aligned}$$

□

3.17 Moment generating functions

Let X be a real random variable. Since, for any $\theta \in \mathbb{R}$, the random variable $e^{\theta X}$ is nonnegative, its expectation exists (but may be equal to $+\infty$). We define the function $M : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ by

$$M(\theta) := \mathbf{E}(e^{\theta X}), \quad \theta \in \mathbb{R}.$$

Notice that $M(0) = 1$. This function is useful if $M(\theta) < \infty$ for some $\theta \neq 0$. (Indeed, there are cases where $\theta = 0$ is the only point at which M is finite.) If X is a positive random variable, then $M(\theta) < \infty$ for all $\theta \leq 0$. If X is a negative random variable, then $M(\theta) < \infty$ for all $\theta \geq 0$. M depends only on the law of X. Indeed, using the Substitution Rule (Lemma 3.23) we can write

$$M(\theta) = \mathbf{E}_{\mathbf{P}_X}(e^{\theta\iota}), \quad \text{where } \iota(x) \equiv x,$$

and, if X has absolutely continuous distribution function F with density f, we can write

$$M(\theta) = \int_{\mathbb{R}} e^{\theta x} f(x) dx \quad \text{(Lebesgue integral)}\ .$$

The function M is called MOMENT GENERATING FUNCTION because of the following:

Lemma 3.32. *Suppose there exist $a < 0 < b$ such that $M(\theta) < \infty$ for all $a < \theta < b$. Then*
(i) the r-moment of X exists for all $r \in \mathbb{N}$ and is given by the r-derivative of M at 0:

$$\mathbf{E}(X^r) = D^r M(0).$$

(ii)

$$M(\theta) = \sum_{r=0}^{\infty} \frac{\mathbf{E}(X^r)}{r!} \theta^r, \quad a < \theta < b.$$

(iii) There is only one distribution function F such that if X has distribution function F then it has moment generating function M.

Proof [sketch] Using the Dominated Convergence Theorem 3.6, we can see that M is infinitely differentiable at 0 with r-derivative equal to the r-moment of X. Moreover, we can see that M is a real analytic function around 0. Hence Taylor's theorem holds, which yields the second claim. □

3.18 APPENDIX A: Sets

The logic of events plays a much more fundamental rôle in the whole subject than the layman could imagine. We review basic concepts here.[h]

§ We express the statement that x is an element of (or belongs to) a set A by $x \in A$. The negation of the latter is denoted by $x \notin A$. We write $A \subset B$ for to express the statement "$x \in A \Rightarrow x \in B$". We let $A \cap B$ be the intersection and $A \cup B$ the union of A and B. To save space, we often write AB in lieu of $A \cap B$ and ABC in lieu of $A \cap B \cap C$. The set A^c is defined through: "$x \in A^c \iff x \notin A$". Naïve set theory and first order logic are algebraically equivalent through the above and through: "$A \cap B \iff x \in A$ and $x \in B$", "$A \cup B \iff x \in A$ or $x \in B$". We let $B \setminus A = B \cap A^c$ and we write $B - A$ for $B \setminus A$ if $A \subset B$. The symmetric difference ('exclusive or') $A \triangle B$ is defined as the set of elements of A or B which do not belong to both, i.e. $A \triangle B = (A \setminus B) \cup (B \setminus A) = (A \cup B) - (A \cap B)$. If $\{A_j, j \in J\}$ is a collection of sets we let $\cap_{j \in J} A_j$ be their intersection and $\cup_{j \in J} A_j$ their union. Here, J is an arbitrary set. If $J = \mathbb{N}$ we write $\cap_{j \in \mathbb{N}} A_j = \cup_{j=1}^{\infty} A_j$, and, as usually in Mathematics, the convention is that the symbol ∞ appearing on top is not part of the index set (∞ is not a natural number).[i] Here is an example of how we translate elementary logic (i.e. ordinary language, say English) in Mathematics: Consider the sentence

$$\mathsf{S} = [\text{after some day it will never rain again in Glasgow}].$$

If we let

$$\mathsf{R}_i = [\text{it rains on the } i\text{-th day in Glasgow}]$$

we can write S as

$$\mathsf{S} = [\exists i \ \forall j \ \geq i \ \text{not}\mathsf{R}_i].$$

Taking the negation of this sentence[j] we get:

$$\text{not}\mathsf{S} = \text{not}[\exists i \ \forall j \ \geq i \ \text{not}R_i] = [\forall i \ \exists j \geq i \ R_i].$$

Translating this back into English, it reads:

$$\text{not}\mathsf{S} = [\text{for all days there will be a day in the future during which it rains in Glasgow}]$$

and, since this is not too palatable even for the literati, we can express it, equivalently, as

$$\text{not}\mathsf{S} = [\text{it rains in Glasgow infinitely often}].$$

What does this have to do with sets? Well, imagine there is a big set (the "universe")–call it Ω–containing all possible states of Glasgow at all days. At the minimum, we want to know whether it rains or not on each day. So we decide to let an element of Ω be the sequence $(\omega_1, \omega_2, \ldots)$ of states of Glasgow, where ω_i takes values "rain" or "doesntrain" and represents what happens on the i-th day in Glasgow. So Ω is the set of all these sequences, mathematically written as $\Omega = \{\text{rain}, \text{doesntrain}\}^{\mathbb{N}}$. Then R_i can be taken to be the set of all sequences $\omega \in \Omega$ such that $\omega_i = \text{rain}$. You can then check that the set (subset of Ω) $S = \cup_i \cap_{j \geq i} R_i^c$ corresponds to the statement S, while $S^c = \cap_i \cup_{j \geq i} R_i$ corresponds to the statement not$\mathsf{S}$. In this sense, the so-called Kolmogorov model of Probability merely puts within a mathematical framework (that of Set Theory) what we intuitively feel as assigning numbers representing chance to statements, to events, to situations describable by ordinary language. So Probability Theory is ordinary logic together with numbers assigned to statements. We talk about logic in this Appendix and about numbers in the next.

[h] There are no numbers in this section; only sets. (Of course, numbers are sets; so the above declaration is offered with apologies to set theorists.)

[i] Compare this with a sum of real numbers: $\sum_{j=1}^{\infty} x_j$ means that j runs over $1, 2, 3 \ldots$ but does not take the value ∞.

[j] To which, I think, we can safely assign the truth value

§ If A is a set then 2^A is the set of its subsets. A set of sets (or "collection" of sets) is a subset of 2^A. If A has n elements then 2^A has 2^n elements. If $\mathbb{N}$ is the set of natural numbers then $2^{\mathbb{N}}$ has as many elements as the set of real numbers. (And why, on earth, should we use the symbol 2^A? To explain, consider a simple set, say $A = \{x, y, z\}$. I can form a subset of A by putting a 1 to each element I pick or 0 to each element I do not pick. For instance, 011 means "do not pick x, pick y, pick z", i.e. it refers to the subset $\{y, z\}$. We thus see there is a correspondence between subsets of A and triples of 0's and 1's. How many such triples do we have? Obviously, $2 \times 2 \times 2 = 2^3$. This explains why we have $2^3 = 8$ subsets. If A had n elements, you see there are 2^n subsets and so the notation 2^A is suggestive of this, and we carry it on even to cases where A has infinite number of elements.)

§ If $f : X \to Y$ is a function then, for any $B \subset Y$ we define $f^{-1}(B) = \{x \in X : f(x) \in B\}$, and for any $A \subset X$ we define $f(A) = \{f(x) : x \in A\}$. Thus f can be lifted as a map from 2^X into 2^Y and we also have a map f^{-1} from 2^Y into 2^X. We always have $f(f^{-1}(B)) \subset B$ and $f^{-1}(f(A)) \supset A$. Also, $f^{-1}(\cap_j B_j) = \cap_j f^{-1}(B_j)$ $f^{-1}(\cap_j B_j) = \cap_j f^{-1}(B_j)$, $f^{-1}(B^c) = f^{-1}(B)^c$. However, f as a map on sets does not behave that well (fortunately!) because $f(\cap A_j) \subset \cap_j f(A_j)$ (equality, in general, fails), but $f(\cup A_j) = \cup_j f(A_j)$; also, $f(A_2 - A_1) \subset f(A_2) - f(A_1)$.

§ The set of functions from X into Y is denoted by Y^X. The notation is motivated by the fact that the set of functions from $\{1, \ldots, n\}$ into $\mathbb{R}$ is "the same" as $\mathbb{R}^n$. Consider, for example, the set of functions from a set A into $\{0, 1\}$. This is denoted as $\{0, 1\}^A$. Since there is a one-to-one correspondence between this set of functions and the collection of subsets 2^A of A (as explained above), the surprise that you may have just experienced with this new notation should not be as high.

§ If $A \subset \Omega$ the function $\mathbf{1}_A : \Omega \to \mathbb{R}$, defined by $\mathbf{1}_A(\omega) = 1$ if $\omega \in A$, and 0, otherwise, is the indicator[k] of the set A. Note that $\mathbf{1}_A \mathbf{1}_B = \mathbf{1}_{AB}$, $\mathbf{1}_{A^c} = 1 - \mathbf{1}_A$. If $A \subset B$, then $\mathbf{1}_{B-A} = \mathbf{1}_B - \mathbf{1}_A$. If $AB = \varnothing$ then $\mathbf{1}_{A \cup B} = \mathbf{1}_A + \mathbf{1}_B$. We can check that the 'inclusion-exclusion formula' $\mathbf{1}_{A \cup B} = \mathbf{1}_A + \mathbf{1}_B - \mathbf{1}_{AB}$ holds for arbitrary A, B, by (if we want to be slick) using $(1 - \mathbf{1}_A)(1 - \mathbf{1}_B) = 1 - \mathbf{1}_A - \mathbf{1}_B + \mathbf{1}_A \mathbf{1}_B$ and using the fact that $\mathbf{1}_A \mathbf{1}_B = \mathbf{1}_{AB}$ while $(1 - \mathbf{1}_A)(1 - \mathbf{1}_B) = \mathbf{1}_{A^c} \mathbf{1}_{B^c} = \mathbf{1}_{A^c B^c} = \mathbf{1}(A \cup B)^c = 1 - \mathbf{1}_{A \cup B}$. The inclusion-exclusion formula for many sets follows by expanding the product $(1 - \mathbf{1}_{A_1}) \cdots (1 - \mathbf{1}_{A_n})$.

§ Often, sets have elements which are themselves sets. (For example, the set of all sets of dishes in a shop.) When we talk about sets of sets we refer to them as classes (or collections, or systems, or aggregates) of sets.[l]

§ A class of sets $\mathscr{R} \subset 2^\Omega$ is a π-system if it is closed under pairwise intersections ($A, B \in \mathscr{R} \Rightarrow AB \in \mathscr{R}$). If $\mathscr{A} \subset 2^\Omega$ then the intersection of all π-systems containing $\mathscr{A}$ is a π-system and is denoted by $\pi(\mathscr{A})$.

§ A class of sets $\mathscr{D} \subset 2^\Omega$ is a λ-system if it is closed under increasing differences ($A, B \in \mathscr{D}, A \subset B \Rightarrow AB \in \mathscr{D}$), and increasing limits ($A_n \in \mathscr{D}, A_n \subset A_{n+1}$, for all n, $\Rightarrow \cup_n \in \mathscr{D}$). If $\mathscr{A} \subset 2^\Omega$ then the intersection of all λ-systems containing $\mathscr{A}$ is a λ-system and is denoted by $\lambda(\mathscr{A})$.

§ A class of sets $\mathscr{C} \subset 2^\Omega$ is a field (or algebra) if it is a π system containing Ω and such that $A \in \mathscr{C} \Rightarrow A^c \in \mathscr{C}$. (Therefore, a field is closed under unions and any finite number of operations on its elements.) If $\mathscr{A} \subset 2^\Omega$ then the intersection of all fields containing $\mathscr{A}$ is a field.

§ A class of sets $\mathscr{F} \subset 2^\Omega$ is a σ-field (or σ-algebra)[m] if it is a field and if $A_1, A_2, \ldots \in \mathscr{F} \Rightarrow \cap_{j=1}^\infty A_j \in \mathscr{F}$. If $\mathscr{A} \subset 2^\Omega$ then the intersection of all λ-systems containing $\mathscr{A}$ is a σ-field and is denoted by $\sigma(\mathscr{A})$.

[k] Also called characteristic function, but, mostly, in Analysis. In Probability, if $\mathbf{1}_A$ is thought of as a random variable, it is also called Bernoulli random variable.

[l] For philological, rather than mathematical, reasons

[m] The letter σ is supposed to mean "sum" and, in fact, a "sum" of countably many objects; in the case at hand it represents "countable unions"

§ If $\{\mathscr{F}_j\}$ is a collection of σ-fields on the same set then $\cap_j \mathscr{F}_j$ is a σ-field too: it is the largest σ-field contained in all of them. (The intersection of σ-fields is never empty because every σ-field contains the empty set.) However the union of σ-fields is not a nice object and we don't consider it. Instead, we define $\vee_j \mathscr{F}_j$ to be the smallest σ-field containing all the $\mathscr{F}_j$.

§ If A, B are sets, then $A \times B$ is the set of all ordered pairs (a, b) where $a \in A$ and $b \in B$. Similarly, if $\{A_j, j \in \mathbb{N}\}$ is a sequence of sets then $\bigtimes_{j=1}^{\infty} A_j$ is the set of all sequences $(a_1, a_2, \ldots)$ where $a_j \in A_j$ for all j. Finally, if $\{A_t, t \in T\}$ is an arbitrary collection of sets we may defined $\bigtimes_{t \in T} A_t$ as the collection of functions $f \in A^T$, where $A = \cup_{t \in T} A_t$ such that $f(t) \in A_t$ for all $t \in T$. Notice that this definition does not respect any "order" in the index set T even if this is there; the order imposed in the definition of the Cartesian product of finitely or countably many sets is an extra bag that we have been forced to carry (for good reasons). When $A = \bigtimes_{j \in J} A_j$ is a Cartesian product, we can define, for each $i \in J$, the projection function $\rho_i : A \to A_i$ by $\rho_i : (a_j, j \in J) \mapsto a_i$.

§ Let $(\Omega_1, \mathscr{F}_1)$, $(\Omega_2, \mathscr{F}_2)$ be measurable spaces. Then $\mathscr{F}_1 \otimes \mathscr{F}_2$ is the σ-field on $\Omega_1 \times \Omega_2$ generated by sets of the form $A_1 \times A_2$, where $A_1 \in \mathscr{F}_1$, $A_2 \in \mathscr{F}_2$. It also equals $\sigma(\rho_1, \rho_2)$, where ρ_1, ρ_2 are the projection functions. It also equals the smallest σ-field containing the σ-field $\{A_1 \times \Omega_2 : \ A_1 \in \mathscr{F}_1\}$ and the σ-field $\{\Omega_2 \times A_2 : \ A_2 \in \mathscr{F}_2\}$. The definition is extended for any finite product.

§ The Borel σ-field $\mathscr{B}$ on $\mathbb{R}$ is the σ-field generated by the class of open sets. (A set O is open if for any any $x \in O$ there is an $\varepsilon > 0$ such that $[x - \varepsilon, x + \varepsilon] \in O$.) It is also generated by intervals of the form $(a, b), a, b \in \mathbb{R}$. It is also generated by intervals of the form $[a, b], a, b \in \mathbb{R}$. It is also generated by intervals of the form $(-\infty, x]$, $x \in \mathbb{R}$. (This last class of intervals forms a π-system.) The Borel σ-field $\mathscr{B}(\mathbb{R}^2)$ on $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$ is defined by $\mathscr{B}(\mathbb{R}^2) = \mathscr{B} \otimes \mathscr{B}$. Similarly, we define $\mathscr{B}(\mathbb{R}^d)$ for any natural number d, as $\mathscr{B}(\mathbb{R}^d) = \mathscr{B}(\mathbb{R}^{d-1}) \otimes \mathscr{B}$, by induction on d. Notice that $\mathscr{B}(\mathbb{R}^d)$ is also generated by open sets in $\mathbb{R}^d$. (A set $O \subset \mathbb{R}^d$ is open if for any $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ there is an $\varepsilon > 0$ such that the cube $[x_1 - \varepsilon, x_1 + \varepsilon] \times \cdots \times [x_d - \varepsilon, x_d + \varepsilon] \in O$.)

§ Let $(S_i, \mathscr{S}_i)$, $i \in \mathbb{N}$ be measurable spaces, and let $S = \bigtimes_i S_i$. A rectangle is a subset of S of the form $B_1 \times \cdots \times B_j \times \cdots$ with $B_i \in \mathscr{S}_i$ for all i. A finite-dimensional rectangle is a rectangle such that $B_k = S_k$ for all k greater than some index j. A cylinder set is a subset of S of the form $A \times S_{j+1} \times S_{j+2} \times \cdots$, where $A \in \mathscr{S}_1 \otimes \cdots \otimes \mathscr{S}_j$. A finite-dimensional rectangle is a rectangle and a cylinder set. A rectangle is not necessarily a cylinder set. The collection of rectangles is π-system (the intersection of two rectangles is still a rectangle; but the union of two rectangles is not a rectangle; the complement of a rectangle is not a rectangle). Ditto for finite-dimensional rectangles. The collection of cylinder sets is a field (the intersection of two cylinder sets is a cylinder set and the complement of a cylinder set is a cylinder set), but not a σ-field. On S we can put a natural σ-field, denoted by $\bigotimes_i \mathscr{S}_i$: it is the σ-field generated by cylinder sets; it is also generated by rectangles; it is also generated by finite-dimensional rectangles; it is also the smallest σ-field such that every projection $\pi_i : (S, \mathscr{S}) \to (S_i, \mathscr{S}_i)$ is a random variable.

§ Let A_n be a sequence of subsets of Ω. We define[n]

$$\overline{\lim_{n\to\infty}} A_n = \cap_n \cup_{m \ge n} A_m, \qquad \underline{\lim_{n\to\infty}} A_n = \cup_n \cap_{m \ge n} A_m,$$

and notice that $(\overline{\lim}_{n\to\infty} A_n)^c = \underline{\lim}_{n\to\infty} A_n^c$. We say that $\lim_{n\to\infty} A_n = A$ (we also write $A_n \to A$) if, by definition, $\overline{\lim}_{n\to\infty} A_n = \underline{\lim}_{n\to\infty} A_n = A$, and write this also as $A_n \to A$. For example, if A_n is an increasing sequence of sets then (show this!) $\lim_{n\to\infty} A_n = \cup_n A_n$ and if A_n is decreasing then $\lim_{n\to\infty} A_n = \cap_n A_n$. Also notice that $\omega \in \overline{\lim} A_n$ if and only if ω belongs to A_n for infinitely many indices n, while $\omega \in \underline{\lim} A_n$ if and only if ω belongs to A_n for all except finitely many n.

[n] Notice the complete analogy with numerical sequences: $\overline{\lim}_{n\to infty} x_n = \inf_n \sup_{k \ge n} x_k$, $\underline{\lim}_{n\to infty} x_n = \sup_n \inf_{k \ge n} x_k$; and $x_n \to x$, as $n \to \infty$, if and only if $\overline{\lim}\, x_n = \underline{\lim}\, x_n = x$; the limit x could be $\pm\infty$.

§ If $\mathscr{S}$ is a λ-system and $A \subset \Omega$ then

$$\mathscr{S}_A := \{B \in 2^\Omega : \ BA \in \mathscr{S}\}$$

is also a λ-system. Indeed, if $B_1 \subset B_2$ both belong to $\mathscr{S}_A$ then B_1A, B_2A belong to $\mathscr{S}$ and $(B_2 - B_1)A = B_2A - B_1A \in \mathscr{S}$, because $\mathscr{S}$ is closed under proper differences; so $B_2 - B_1 \in \mathscr{S}_A$. Also, if $B_n \in \mathscr{S}_A$, $B_n \subset B_{n+1}$ for all n, then $B_nA \in \mathscr{S}$ for all n, and so $A \cup_n B_n = \cup_n AB_n \in \mathscr{S}$ because $\mathscr{S}$ is closed under increasing limits; so $\sup_n B_n \in \mathscr{S}_A$.

§ π & λ & $\Omega \equiv \sigma$. In other words, if $\mathscr{S}$ is both π and λ and contains Ω then it is a σ-field. Indeed, if $A \in \mathscr{S}$ then $A^c = \Omega - A \in \mathscr{S}$ because $\mathscr{S}$ is λ. So $\mathscr{S}$ is closed under complementations, intersections and limits. So it is a σ-field.

§**[Sierpiński-Dynkin Lemma]** Let $\mathscr{S}$ be a π-system containing Ω. Then $\lambda(\mathscr{S}) = \sigma(\mathscr{S})$.

Proof Since $\lambda(\mathscr{S})$ is the smallest λ-system containing $\mathscr{S}$ and since $\sigma(\mathscr{S})$ is a λ-system, we have $\lambda(\mathscr{S}) \subset \sigma(\mathscr{S})$. We need to show that $\sigma(\mathscr{S}) \subset \lambda(\mathscr{S})$. If we show that $\lambda(\mathscr{S})$ is also a π-system then it will be a σ-field and we'll be done. We know that

$$A \in \mathscr{S}, \ B \in \mathscr{S} \Rightarrow AB \in \mathscr{S} \subset \lambda(\mathscr{S}). \tag{3.12}$$

Hence

$$\forall A \in \mathscr{S} \quad \mathscr{S} \subset \{B \in 2^\Omega : AB \in \lambda(\mathscr{S})\} =: \lambda(\mathscr{S})_A.$$

But $\lambda(\mathscr{S})_A$ is λ and so

$$\forall A \in \mathscr{S} \quad \lambda(\mathscr{S}) \subset \lambda(\mathscr{S})_A.$$

Let's read this again:

$$A \in \mathscr{S}, \ B \in \lambda(\mathscr{S}) \Rightarrow AB \in \lambda(\mathscr{S}). \tag{3.13}$$

Wonder of wonders: (3.12) implied the seemingly stronger (3.13). Let's rewrite the (3.13) as:

$$\forall B \in \lambda(\mathscr{S}) \quad \mathscr{S} \subset \{A \in 2^\Omega : \ AB \in \lambda(\mathscr{S})\} =: \lambda(\mathscr{S})_B.$$

But $\lambda(\mathscr{S})_B$ is λ and so

$$\forall B \in \lambda(\mathscr{S}) \quad \lambda(\mathscr{S}) \subset \lambda(\mathscr{S})_B.$$

Let's read this again:

$$B \in \lambda(\mathscr{S}), \ A \in \lambda(\mathscr{S}) \Rightarrow AB \in \lambda(\mathscr{S}). \tag{3.14}$$

Wonder of wonders of wonders: (3.13) implied the seemingly stronger (3.14) which is what we need. □

3.19 APPENDIX B: Sets and numbers

In some sense, probability is all about assigning numbers to sets.

§ Let Ω be a set and $\mathscr{C}$ a field of subsets of it. The function $\mathbf{P}_0 : \mathscr{C} \to [0,1]$ is additive if $\mathbf{P}_0(A \cup B) = \mathbf{P}_0(A) + \mathbf{P}_0(B)$ whenever $AB = \varnothing$. It is countably additive if for any $A \in \mathscr{C}$ for which there exist mutually disjoint $A_n \in \mathscr{C}$ with $A = \cup_n A_n$ we have $\mathbf{P}_0(A) = \sum_n \mathbf{P}_0(A_n)$. We will throughout assume that $\mathbf{P}_0(\Omega) = 1$. (If, in addition to the above, we have that $\mathscr{C}$ is a σ-field, then $\mathbf{P}_0$ is a probability.)

§ If $\mathbf{P}$ is a probability on $(\Omega, \mathscr{F})$, then $(\Omega, \mathscr{F}, \mathbf{P})$ is called probability space. If A_n is a sequence of sets in $\mathscr{F}$ then $\mathbf{P}(\cup_n A_n) \le \sum_n \mathbf{P}(A_n)$ [Boole's inequality]. If $A_n \to A$ as $n \to \infty$ then $\mathbf{P}(A_n) \to \mathbf{P}(A)$ as $n \to \infty$ [continuity]. If $\sum_n \mathbf{P}(A_n) < \infty$ then $\mathbf{P}(\overline{\lim} A_n) = 0$. [Borel-Cantelli lemma]. (Proof: $\cup_{k \ge m} A_k \to \overline{\lim} A_n$, as $m \to \infty$; So $\mathbf{P}(\overline{\lim}_n A_n) = \lim_{k \ge m} \mathbf{P}(\cup_{k \ge m} A_k)$. But the latter is $\le \sum_{k \ge m} \mathbf{P}(A_k)$ which tends to 0, as $m \to \infty$, by the assumption.)

Lemma [CA] The following are equivalent:
(i) $\mathbf{P}_0$ is countably additive on the field $\mathscr{C}$.
(ii) $\mathbf{P}_0(A_n) \to 0$ for all decreasing sequences $A_n \in \mathscr{C}$ with $\cap_n A_n = \varnothing$.
(iii) $\mathbf{P}_0(A_n) \to 0$ for all sequences $A_n \in \mathscr{C}$ with $A_n \to \varnothing$.

Proof Suppose (i) holds. Let A_n decrease to $\varnothing$. We have

$$\begin{aligned}\mathbf{P}_0(A_1) &= \sum_{k=1}^{\infty} \mathbf{P}_0(A_k - A_{k+1}) \\ &= \lim_{n\to\infty} \sum_{k=1}^{n-1} [\mathbf{P}_0(A_k) - \mathbf{P}_0(A_{k+1})] \\ &= \lim_{n\to\infty} [\mathbf{P}_0(A_1) - \mathbf{P}_0(A_n)],\end{aligned}$$

i.e. $\mathbf{P}_0(A_n) \to 0$. We proved (i) ⇒ (ii). Next, suppose (ii) holds. Let A_n be mutually disjoint elements of $\mathscr{C}$ with $\cup_n A_n \in \mathscr{C}$. By finite additivity,

$$\mathbf{P}_0(\cup_k A_k) = \sum_{k=1}^{n} \mathbf{P}_0(A_k) + \mathbf{P}_0(\cup_{k>n} A_k).$$

Since $\cup_{k>n} A_k$ decreases to $\varnothing$ we have $\mathbf{P}_0(\cup_{k>n} A_k) \to 0$, and so we obtain countable additivity. We proved (ii) ⇒ (i). Suppose (ii) holds. Let $A_n \in \mathscr{C}$, $A_n \to \varnothing$. We want to show that $\mathbf{P}_0(A_n) \to 0$. The problem is that A_n is not decreasing so we need, somehow, to construct auxiliary decreasing sequence of sets. What does $A_n \to 0$ mean? It means that

$$G_n := \cup_{k \geq n} A_k \downarrow \varnothing.$$

If we knew that $G_n \in \mathscr{C}$ we would immediately get the result. The problem is we don't know that, so we can't use G_n as an argument of the function $\mathbf{P}_0$. We need to work harder. We only know that finite unions belong to $\mathscr{C}$. So let us approximate G_n. We have

$$G_{n,r} := \cup_{k=n}^{r} A_k \uparrow G_n, \quad \text{as } r \to \infty.$$

We also know that $G_{n,r} \in \mathscr{C}$ for all n, r. So we can talk about $\mathbf{P}_0(G_{n,r})$, a numerical sequence which increases with r (and bounded below 1), having a limit, say

$$\gamma_n = \lim_{r\to\infty} \mathbf{P}_0(G_{n,r}).$$

This means that, FOR EACH n, $\mathbf{P}_0(G_{n,r})$ differs from γ_n by anything we like, when r is large. Let ε_n be this "anything we like" thing. We just said that

$$\text{FOR ALL } n, \quad \gamma_n - \mathbf{P}_0(G_{n,r}) \leq \varepsilon_n, \quad \text{for all } r \text{ larger than or equal to some } r(n).$$

In particular,

$$\text{FOR ALL } n, \quad \gamma_n - \mathbf{P}_0(G_{n,r(n)}) \leq \varepsilon_n.$$

And if we let

$$R(n) := \max(r(1), r(2), \ldots, r(n)),$$

we have

$$\text{FOR ALL } n, \quad \gamma_n - \mathbf{P}_0(G_{n,R(n)}) \leq \varepsilon_n,$$

because $r(n) \leq R(n)$ and so $\mathbf{P}_0(G_{n,R(n)}) \leq \mathbf{P}_0(G_{n,r(n)})$. Moreover, the sequence $R(n)$ is increasing. We have

$$A_n \subset G_{n,R(n)} \subset G_n \downarrow \varnothing,$$

so, while we have created a sequence $G_{n,R(n)}$ of sets in $\mathscr{C}$ which is above A_n, and does converge to $\varnothing$, we do not know that $G_{n,R(n)}$ decreases, so we cannot apply our assumption (ii). Therefore we need to work harder. We take

$$H_n = \cap_{k=1}^{n} G_{k,R(k)}.$$

Clearly, $H_n \in \mathscr{C}$, and H_n decreases, and $H_n \downarrow \varnothing$, therefore, by (ii),

$$\mathbf{P}_0(H_n) \to 0.$$

But how does H_n compare to A_n? We lost track of it, so let's see:

$$\begin{aligned}
\mathbf{P}_0(A_n \setminus H_n) &= \mathbf{P}_0\left(\cup_{k=1}^n (A_n \setminus G_{k,R(k)})\right) \\
&\le \sum_{k=1}^n \mathbf{P}_0\left(A_n \setminus G_{k,R(k)}\right) \\
&\le \sum_{k=1}^n \mathbf{P}_0\left(G_{n,R(n)} - G_{k,R(k)}\right) = \sum_{k=1}^n \left(\mathbf{P}_0(G_{n,R(n)}) - \mathbf{P}_0(G_{k,R(k)})\right) \\
&\le \sum_{k=1}^n \left(\mathbf{P}_0(G_{k,R(n)}) - \mathbf{P}_0(G_{k,R(k)})\right) \\
&\le \sum_{k=1}^n \left(v_k - \mathbf{P}_0(G_{k,R(k)})\right) \le \sum_{k=1}^n \varepsilon_k,
\end{aligned}$$

where, recall, ε_k is ANYTHING WE PLEASE. We would be pleased to show that $\mathbf{P}_0(A_n \setminus H_n) \to 0$, for this would imply $\mathbf{P}_0(A_n) \to 0$. Let us then CHOOSE

$$\varepsilon_k = \eta 2^{-k},$$

where $\eta > 0$ is arbitrary. Hence $\sum_{k=1}^n \varepsilon_k \le \eta$. Hence $\mathbf{P}_0(A_n \setminus H_n) \le \eta$, and so we can conclude. We proved that (ii) $\Rightarrow$ (iii). Since, obviously, (iii) $\Rightarrow$ (ii), the lemma has been proved. □

Lemma [COMMON LIMITS] Suppose that $\mathbf{P}_0$ is countably additive on the field $\mathscr{C}$.
(i) If A_n is a sequence in $\mathscr{C}$ that has a limit (but which is not necessarily in $\mathscr{C}$, then $\mathbf{P}_0(A_n)$ has a limit.
(ii) If A_n, B_n are sequences in $\mathscr{C}$ with the same limit (not necessarily in $\mathscr{C}$) then $\mathbf{P}_0(A_n)$, $\mathbf{P}_0(B_n)$ have the same limit.

Proof Recall that a sequence x_n of real numbers converges if and only if, given $\varepsilon > 0$, there is an integer N such that $\sup_{m>n} |x_n - x_m| \le \varepsilon$ if $n > N$. The negation of the latter statement implies the existence of $\varepsilon > 0$ such that, for all N there is $n > N$ and $m > N$ with $|x_n - x_m| > \varepsilon$; or, if we let $m = I(n)$, we have that $|x_n - x_{I(n)}| \not\to 0$, as $n \to \infty$. Therefore, if we can prove that for <u>all</u> strictly increasing sequence of integers $I(n)$ we have

$$x_n - x_{I(n)} \to 0, \quad \text{as } n \to \infty,$$

we can deduce that x_n has a limit. Let $I(n)$ be such a sequence. Then

$$|\mathbf{P}_0(A_n) - \mathbf{P}_0(A_{I(n)}| \le |\mathbf{P}_0(A_n \triangle A_{I(n)}|.$$

But A_n has a limit. Therefore $A_n \triangle A_{I(n)} \to \varnothing$. By the Lemma above, $\mathbf{P}_0(A_{I(n)} \to 0$, and so (i) Is proved. To prove (ii), we use the same reasoning but with $B_{I(n)}$ instead of $A_{I(n)}$. □

Lemma [FIRST EXTENSION] Suppose that $\mathbf{P}_0$ is countably additive on the field $\mathscr{C}$. Let $D(\mathscr{C})$ be the class of subsets of Ω that are limits of sets in $\mathscr{C}$. Define $\mathbf{P}_1$ on $D(\mathscr{C})$ by

$$\mathbf{P}_1(\lim_n A_n) = \lim_n \mathbf{P}_0(A_n), \quad A_n \in \mathscr{C}.$$

Then
(i) $D(\mathscr{C})$ is a field.
(ii) The definition $\mathbf{P}_1$ is valid.
(iii) $\mathbf{P}_1$ is countably additive on $D(\mathscr{C})$.

Proof (i) Easy.
(ii) Follows from Lemma [COMMON LIMITS].
(iii) From Lemma [CA], it suffices to show that if A_n is a sequence of elements of $D(\mathscr{C})$ such that $A_n \downarrow \varnothing$ then $\mathbf{P}_1(A_n) \to 0$. Since $A_m \in D(\mathscr{C})$ we can write

$$A_m = \lim_n B_{m,n}.$$

Then we also have

$$A_m = \lim_n (B_{1,n} \cap \cdots \cap B_{m,n}).$$

Choose a sequence of integers n_m increasing to ∞ such that

$$\lim_m [\mathbf{P}_1(A_m) - \mathbf{P}_0(B_{1,n_m} \cap \cdots \cap B_{m,n_m})] = 0.$$

Since

$$\lim_m (B_{1,n_m} \cap \cdots \cap B_{m,n_m}) = \varnothing,$$

we have

$$\lim_m \mathbf{P}_0(B_{1,n_m} \cap \cdots \cap B_{m,n_m}) = 0,$$

and so

$$\lim_m \mathbf{P}_1(A_m) = 0.$$

□

Lemma [SECOND EXTENSION] Suppose that $\mathbf{P}_0$ is countably additive on the field $\mathscr{C}$. Let $D(\mathscr{C})$ be the class of subsets of Ω that are limits of sets in $\mathscr{C}$. Let $D^2(\mathscr{C})$ be the class of subsets of Ω that are limits of sets in $D(\mathscr{C})$. Define $\mathbf{P}_2$ on $D^2(\mathscr{C})$ by

$$\mathbf{P}_2(\lim_n A_n) = \lim_n \mathbf{P}_1(A_n), \quad A_n \in D(\mathscr{C}).$$

Then (i) $D^2(\mathscr{C})$ is a field.
(ii) The definition $\mathbf{P}_2$ is valid.
(iii) $\mathbf{P}_2$ is countably additive on $D^2(\mathscr{C})$.

Proof Repetition of the proof of Lemma [FIRST EXTENSION]. □

Proposition [SANDWICH] Let $\mathscr{F}$ be the σ-field generated by $\mathscr{C}$. Then, for any $B \in \mathscr{F}$, there are $A, C \in D^2(\mathscr{C})$ such that $\mathbf{P}_2(A) = \mathbf{P}_2(C)$.

Proof Let $D^{\downarrow}(\mathscr{C})$ (respectively, $D^{\uparrow}(\mathscr{C})$) be the class of sets which are limits of decreasing (respectively, increasing) sequences of sets of $\mathscr{C}$. Define the class $\mathscr{D}$ of subsets of Ω for which, given $\varepsilon > 0$, there is $G \in D^{\downarrow}(\mathscr{C})$ and $H \in D^{\uparrow}(\mathscr{C})$ such that

$$G \subset A \subset H, \quad \mathbf{P}_1(H) - \mathbf{P}_1(G) < \varepsilon.$$

We have $\mathscr{C} \subset \mathscr{D}$. Also, $\Omega \in \mathscr{D}$. So if we show that $\mathscr{D}$ is both π and λ, then it is σ and so $\sigma(\mathscr{C}) \subset \mathscr{D}$ and we are done. That it is closed under intersections and proper differences is easy. That it is closed under increasing limits is left as an exercise. □

SMSTC (2007/08)

Probability

4: Joint distributions and independence

Takis Konstantopoulos[a]

www.smstc.ac.uk

Contents

4.1 Introduction

Most of this chapter is concerned with a random variable (X, Y) with values in $\mathbb{R}^2$, but the concepts/results are easily generalisable to random variables with values in $\mathbb{R}^d$.

We first explain why the joint distribution function defines the law of (X, Y). We talk about independence and carefully explain why independence between classes of sets closed under intersection implies independence between their σ-fields.

To define conditional density we define projection and prove everything, except the existence of a regular conditional distribution.

Finally, we talk about a Gaussian variable in $\mathbb{R}^d$.

4.2 Joint distributions

Consider a random variable $(X, Y) : (\Omega, \mathscr{F}) \to (\mathbb{R}^2, \mathscr{B}(\mathbb{R}^2))$. This random variable is called TWO RANDOM VARIABLES.

If $\mathbf{P}$ is a probability of $(\Omega, \mathscr{F})$, the joint distribution of (X, Y) is another name for the law $\mathbf{P}_{X,Y}$ of the random variable (X, Y).

The joint distribution function of (X, Y) is the function

$$F_{X,Y}(x, y) := \mathbf{P}(X \le x, Y \le y) = \mathbf{P}_{X,Y}((-\infty, x] \times (-\infty, y]), \quad (x, y) \in \mathbb{R}^2. \tag{4.1}$$

The law $\mathbf{P}_X$ of X is referred to as the first marginal of the law $\mathbf{P}_{X,Y}$. The distribution function F_X of X is referred to as the first marginal distribution function of the joint distribution function $F_{X,Y}$ and, of course,

$$F_X(x) = \lim_{y \to \infty} F_{X,Y}(x, y).$$

Note that we chose ti use $\le$ instead of $<$ in (4.1) for no good reason other than a mere arbitrary convention.

4.2.1 Knowledge of $F_{X,Y}$ implies knowledge of $\mathbf{P}_{X,Y}$

We would like to explain why, knowledge of the function $F_{X,Y}$ implies knowledge of $\mathbf{P}_{X,Y}(B)$ for all $B \in \mathscr{B}(\mathbb{R}^2)$. Consider a rectangle (with sides parallel to the axes–please think geometrically)

$$(a_1, b_1] \times (a_2, b_2] := \{(x, y) \in \mathbb{R}^2 : \ a_1 < x \le b_1, \ a_2 < y \le b_2\}. \tag{4.2}$$

We allow a_1, a_2 to take any value, including $-\infty$. Since $(-\infty, b_1] \times (-\infty, b_2]$ is the disjoint union of four rectangles, using additivity, we obtain

$$\mathbf{P}_{X,Y}((a_1, b_1] \times (a_2, b_2]) = F_{X,Y}(b_1, b_2) - F_{X,Y}(b_1, a_2) - F_{X,Y}(a_1, b_2) + F_{X,Y}(b_1, b_2).$$

Rectangles of the form (4.2) have the following nice properties: First, intersection of two of them is a rectangle of the same form. Thus, if $\mathscr{R}$ denotes the collection of these rectangles we have that $\mathscr{R}$ is closed under finite intersection, i.e. it is a π-system. Second, the complement of a rectangle from $\mathscr{R}$ is a finite union of disjoint rectangles from $\mathscr{R}$.

EXERCISE 31. Write, explicitly, the complement of $(a_1, b_1] \times (a_2, b_2]$ as the disjoint union of elements of $\mathscr{R}$. Notice we can do that in at least two different ways. Do so.

Now consider the class

$$\mathscr{C} := \{\text{finite disjoint unions of elements of } \mathscr{R}\}.$$

It is easy to visualise, geometrically, what kind of elements $\mathscr{C}$ contains. Then

EXERCISE 32. Show that $\mathscr{C}$ is a field, i.e. if $A \in \mathscr{C}$ then $A^c \in \mathscr{C}$ and if $A, B \in \mathscr{C}$ then $A \cup B \in \mathscr{C}$.

Note that the notation used here mimics that of Section 3.7. Indeed, the classes $\mathscr{R}, \mathscr{C}$ here behave in the same way as the classes with the same names there.

Hence, if $A \in \mathscr{C}$, we can write $A = \cup_{i=1}^m B_i$, where B_i are disjoint rectangles from $\mathscr{R}$, and, since for each such rectangle B_i we can use $F_{X,Y}$ to compute $\mathbf{P}_{X,Y}(B)$, we have that

$$\mathbf{P}_{X,Y}(A) = \sum_{i=1}^{m} \mathbf{P}_{X,Y}(B), \quad A \in \mathscr{C},$$

can be computed by using $F_{X,Y}$ only.

But there are many sets in $\mathscr{B}(\mathbb{R}^2)$ that do not belong to $\mathscr{C}$, so we wish to continue our endeavour. If it were true that every set in $\mathscr{B}(\mathbb{R}^2)$ was a limit of sets of $\mathscr{C}$ then we would be finished, by the continuity property of a probability. However, there are many elements in $\mathscr{B}(\mathbb{R}^2)$ that are not limits of sets in $\mathscr{C}$.

Example 4.1. The set of all points (x, y) where x, y are rationals cannot be obtained as a limit of elements of $\mathscr{C}^2$. To illuminate this point, consider "straightforward" procedure that places a little rectangle around each such point and then let the little rectangle shrink. Specifically, let $\mathbb{Q} = \{q_1, q_2, \ldots\}$ be an enumeration of the rationals. To each (q_m, q_n) associate the rectangle

$$I_{m,n}(\varepsilon) := (q_m - \varepsilon 2^{-m},\ q_m + \varepsilon 2^{-m}] \times (q_n - \varepsilon 2^{-n},\ q_n + \varepsilon 2^{-n}],$$

and let $I(\varepsilon) := \cup_{m,n} I_{m,n}(\varepsilon)$. Show that $\cap_{\varepsilon>0, \varepsilon\in\mathbb{Q}} I(\varepsilon)$ is not equal to $\mathbb{Q} \times \mathbb{Q}$.
Hint: The set $I(\varepsilon)$ is uncountable.

We will denote by $\mathbf{P}_0$ the function $\mathbf{P}_{X,Y}$ restricted to $\mathscr{C}$. Clearly, $\mathbf{P}_0$ is uniquely specified by $F_{X,Y}$. If we show that $\mathbf{P}_0$ can be uniquely extended to a probability on $(\mathscr{R}^2, \mathscr{B}(\mathbb{R}^2))$ then our claim that $F_{X,Y}$ completely specifies $\mathbf{P}_{X,Y}$ will be proved. The steps in showing this are precisely the same as those of Section 3.7.

1. We first show that $\mathbf{P}_0$ is countably additive on $\mathscr{C}$.
2. We then extend $\mathbf{P}_0$ to $\mathbf{P}_1$ on $D(\mathscr{C})$.
3. We then extend $\mathbf{P}_1$ to $\mathbf{P}_2$ on $D^2(\mathscr{C})$.
4. Finally, we show, precisely as before, that, for every $B \in \mathscr{B}(\mathscr{R}^2)$ there are sets $A, C \in D^2(\mathscr{C})$ such that $A \subset B \subset C$ and $\mathbf{P}_2(A) = \mathbf{P}_2(C)$.

4.3 Independence

Recall (see $\boxed{2.2}$) that X, Y are independent random variables on $(\Omega, \mathscr{F}\mathbf{P})$ if $\sigma(X)$, $\sigma(Y)$ are independent σ-fields. We wish to show that

Proposition 4.1. *X, Y are independent on $(\Omega, \mathscr{F}, \mathbf{P})$ if and only if $F_{X,Y}(x, y) := \mathbf{P}(X \le x, Y \le Y)$, $F_X(x) := \mathbf{P}(X \le x)$, $F_Y(y) := \mathbf{P}(Y \le y)$ are related by*

$$F_{X,Y}(x, y) = F_X(x)F_Y(y).$$

This is a consequence of the following:

Lemma 4.1. *Let $(\Omega, \mathscr{F}, \mathbf{P})$ be a probability space. Let $\mathscr{R}_1, \mathscr{R}_2$ be two π-systems. Then $\sigma(\mathscr{R}_1), \sigma(\mathscr{R}_2)$ are independent if and only if*

$$\mathbf{P}(B_1B_2) = \mathbf{P}(B_1)\mathbf{P}(B_2), \quad B_1 \in \mathscr{R}_1,\ B_2 \in \mathscr{R}_2.$$

Proof The proof mimics the uniqueness part of the proof of Theorem 3.1. Fix $B_1 \in \mathscr{R}_1$ and consider two probabilities on $\sigma(\mathscr{R}_2)$:

$$\mathbf{P}(B_1A_2), \quad \mathbf{P}(B_1)\mathbf{P}(A_2), \quad A_2 \in \sigma(\mathscr{R}_2).$$

Our assumption says that these two probabilities agree on $\mathscr{R}_2$. Since $\mathscr{R}_2$ is a π-system, the two probabilities agree on $\sigma(\mathscr{R}_2)$: For each $B_1 \in \mathscr{R}_1$,

$$\mathbf{P}(B_1A_2) = \mathbf{P}(B_1)\mathbf{P}(A_2), \quad A_2 \in \sigma(\mathscr{R}_2). \tag{4.3}$$

Now, for fixed $A_2 \in \sigma(\mathscr{R}_2)$ consider two probabilities on $\sigma(\mathscr{R}_1)$:

$$\mathbf{P}(A_1A_2), \quad \mathbf{P}(A_1)\mathbf{P}(A_2), \quad A_1 \in \sigma(\mathscr{R}_1).$$

(4.3) says that these two probabilities agree on $\mathscr{R}_1$. Since $\mathscr{R}_1$ is a π-system, the two probabilities agree on $\sigma(\mathscr{R}_1)$: For each $A_2 \in \sigma(\mathscr{R}_2)$,

$$\mathbf{P}(A_1A_2) = \mathbf{P}(A_1)\mathbf{P}(A_2), \quad A_1 \in \sigma(\mathscr{R}_1),$$

which is what we need. □

Proof of Proposition 4.1: Let $\mathscr{R}_1 = \{X^{-1}(a, b] : \ a < b\}$, $\mathscr{R}_2 = \{Y^{-1}(a, b] : \ a < b\}$. These are both π-systems. It is easy to see that the assumption $F_{X,Y}(x, y) = F_X(x)F_Y(y)$ implies that $\mathbf{P}(B_1B_2) = \mathbf{P}(B_1)\mathbf{P}(B_2)$, $B_1 \in \mathscr{R}_1$, $B_2 \in \mathscr{R}_2$. Hence $\mathbf{P}(A_1A_2) = \mathbf{P}(A_1)\mathbf{P}(A_2)$, $A_1 \in \sigma(\mathscr{R}_1) = \sigma(X)$, $A_2 \in \sigma(\mathscr{R}_2) = \sigma(Y)$. □

4.4 Joint density

Consider the coin flip space $(\Omega = \{0,1\}^{\mathbb{N}}, \mathscr{F}, \mathbf{P})$. Define the random variables

$$U_1 := \sum_{n=1}^{\infty} \frac{\omega_{2n}}{2^n}, \quad U_2 := \sum_{n=1}^{\infty} \frac{\omega_{2n-1}}{2^n}.$$

By applying Lemma 4.1 we can see that U_1, U_2 are independent.

Let $h : (\mathbb{R}^2, \mathscr{B}(\mathbb{R}^2)) \to (\mathbb{R}, \mathscr{B})$ be measurable and suppose $h \geq 0$. Define the Lebesgue integral

$$\int_{\mathbb{R}^2} h(x,y) dx dy := \sum_{n_1, n_2 \in \mathbb{Z}} \mathbf{E} h(U_1 + n_1, U_2 + n_2),$$

and, more generally, for any $B \in \mathscr{B}(\mathbb{R}^2)$,

$$\int_B h(x,y) dx dy := \int_{\mathbb{R}^2} h(x,y) \mathbf{1}_B(x,y) dx dy.$$

If h has no restriction on sign, define $\int_B h(x,y) dx dy := \int_B h^+(x,y) dx dy - \int_B h^-(x,y) dx dy$, provided not both terms are infinite. The following theorem connects what one learns in basic Calculus of Many Variables with what we just defined.

Theorem 4.1. *(i) If h is Riemann integrable on a rectangle $R = [a,b] \times [c,d]$ then its Lebesgue integral on R coincides with its Riemann integral on R.*
(ii) If h is bounded and measurable then it is Riemann integrable on R of the set of discontinuities D of h satisfies $\mathbf{P}((U_1, U_2) \in D) = 0$.

More generally, given two distribution functions F_1, F_2 on $\mathbb{R}$, we can define the product measure $F_1 \times F_2$ on $(\mathbb{R}^2, \mathscr{B}(\mathbb{R}^2))$ by defining it first on rectangles,

$$(F_1 \times F_2)(B_1 \times B_2) := \mathbf{P}(F_1^{-1}(U_1) \in B_1, \ F_2^{-1}(U_2) \in B_2),$$

and then extending it to $\mathscr{B}(\mathbb{R}^2)$ using the procedure explained earlier. We can also define the Lebesgue-Stieltjes integral

$$\int_{\mathbb{R}^2} h \ d(F_1 \times F_2) := \mathbf{E} h(F_1^{-1}(U_1), F_2^{-1}(U_2)).$$

Theorem 4.2 (Fubini)**.** *If $h \geq 0$ or if $\int_{\mathbb{R}^2} |h| \ d(F_1 \times F_2) < \infty$, then*

$$\int_{\mathbb{R}^2} h \ d(F_1 \times F_2) = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} h(x,y) F_2(dy) \right) F_1(dx) = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} h(x,y) F_1(dx) \right) F_2(dy).$$

Proof [sketch]: The statement is obvious when $h(x,y) = \mathbf{1}_{B_1}(x) \mathbf{1}_{B_2}(y)$. Hence it is obvious for linear combinations of such functions. The general statement follows from an approximation procedure relying on an argument based on the Sierpiński-Dynkin lemma. □

Occasionally, it so happens that there exists a function f such that $F_{X,Y}$ can be written as a Lebesgue integral:

$$F_{X,Y}(x,y) = \int_{(-\infty,x] \times (-\infty,y]} f(s,t) ds dt.$$

In such a case, we say that (X,Y) is absolutely continuous (jointly absolutely continuous, I suppose, if you want to be pedantic) and that $f_{X,Y}$ is a density. Using Fubini's theorem, we see that

$$f_X(x) = \int_{(-\infty,x]} f(s) ds$$

is a density for X (i.e. X is absolutely continuous, and so is Y).

Another consequence of Fubini's theorem is:

Lemma 4.2. *If X, Y are independent then*

$$\mathbf{E}(XY) = (\mathbf{E}X)(\mathbf{E}Y),$$

whenever the expectations are defined.

And another, useful consequence of Fubini's theorem is:

Lemma 4.3. *If X is a positive random variable then*

$$\mathbf{E}X = \int_0^\infty \mathbf{P}(X > x)\, dx.$$

A standard criterion for independence between X, Y, for absolutely continuous (X, Y), is:

Lemma 4.4. *Suppose that (X, Y) is absolutely continuous. Let $f_{X,Y}$ be a density of (X, Y). Let f_X, f_Y be densities of X, Y, respectively. Then X, Y are independent if and only if*

$$f(x, y) = f_X(x) f_Y(y),$$

for all (x, y) except, possibly, on a set of measure zero.

4.5 Joint moment generating function

When (X, Y) is a random variable in $\mathbb{R}^2$ we can define its moment generating function by

$$M_{X,Y}(\eta, \theta) := \mathbf{E}e^{\eta X + \theta Y}, \quad \eta, \theta \in \mathbb{R}.$$

whenever it exists. Let also M_X, M_Y be the moment generating functions of X, Y. One can prove that:

Lemma 4.5. *Suppose that $M_{X,Y}$ exists in a (one sided) neighbourhood of zero. If $M_{X,Y}(\eta, \theta) = M_X(\eta) M_Y(\theta)$ then X, Y are independent.*

4.6 Correlations

We now consider random variables X on some probability space $(\Omega, \mathscr{F}, \mathbf{P})$ with $\mathbf{E}X^2 < \infty$. The aggregate of all these random variables will be denoted by

$$L^2(\Omega, \mathscr{F}, \mathbf{P}).$$

If $\mathbf{E}X^2 < \infty$, $\mathbf{E}Y^2 < \infty$, then $\mathbf{E}(X + Y)^2 < \infty$ (indeed, $(x + y)^2 \leq 2x^2 + 2y^2$). This means that if $X, Y \in L^2(\Omega, \mathscr{F}, \mathbf{P})$ then, for any $a, b \in \mathbb{R}$, $aX + bY \in L^2(\Omega, \mathscr{F}, \mathbf{P})$. Hence $L^2(\Omega, \mathscr{F}, \mathbf{P})$ is a linear space (a linear subspace of $\mathbb{R}^\Omega$). Recall that the covariance between X and Y is defined by

$$\operatorname{cov}(X, Y) = \mathbf{E}(X - \mathbf{E}X)(Y - \mathbf{E}Y).$$

Since, for $X_1, X_2, Y \in L^2(\Omega, \mathscr{F}, \mathbf{P})$, $a_1, a_2 \in \mathbb{R}$,

$$\operatorname{cov}(a_1 X_1 + a_2 X_2, Y) = a_1 \operatorname{cov}(X_1, Y) + a_2 \operatorname{cov}(X_2, Y),$$

the covariance is linear in each of its arguments when the other is kept fixed and it can thus be used to define an inner product:

$$\langle X, Y \rangle := \operatorname{cov}(X, Y).$$

We also define the semi-norm

$$||X|| := \sqrt{\operatorname{cov}(X, X)} = \sqrt{\operatorname{var}(X, X)},$$

where the word 'semi-norm' means that it has the following properties:

1. $||X|| \geq 0$
2. $||a_1 X_1 + a_2 X_2|| = |a_1| \; ||X_1|| + |a_2| \; ||X_2||$.
3. $||X + Y|| \leq ||X|| + ||Y||$.

Another name for $||X||$ is 'standard deviation'. We also note that

If $||X|| = 0$ then $\mathbf{P}(X = 0) = 1$.

We cannot deduce, from $||X|| = 0$ alone that $X(\omega) = 0$ for all $\omega \in \Omega$, but only that $X(\omega) = 0$ for all ω except those in a set of probability zero. If we could deduce that $X(\omega) = 0$ for all $\omega \in \Omega$, we would say that $|| \cdot ||$ is a norm. To get around this problem, we merely identify all random variables in $L^2(\Omega, \mathscr{F}, \mathbf{P})$ which differ on a set of measure zero: That is, we let $[X]$ be the set of all Y such that $\mathbf{P}(X \neq Y) = 0$, and redefine $L^2(\Omega, \mathscr{F}, \mathbf{P})$ to be the collection of all such $[X]$. It is not hard to see that this is still a linear space and if we let $||[X]|| := ||X||$ (which is well defined), then this is a norm.

Being a normed space with an inner product, $L^2(\Omega, \mathscr{F}, \mathbf{P})$ has a structure much like the geometric structure of the usual Euclidean space, for instance, Pythagoras' theorem holds:

$$||X + Y||^2 = ||X||^2 + ||Y||^2 \quad \text{if } \langle X, Y \rangle = 0.$$

The notion of convergence in $L^2(\Omega, \mathscr{F}, \mathbf{P})$ is as follows: We say that $X_n \to X$ if $||X_n - X|| \to 0$, as $n \to \infty$. With respect to this notion of convergence, the space $L^2(\Omega, \mathscr{F}, \mathbf{P})$ is complete:

Lemma 4.6. *If X_n is a sequence in $L^2(\Omega, \mathscr{F}, \mathbf{P})$ such that*

$$\sup_{m,n \geq N} ||X_n - X_m|| \to 0 \textit{ as } N \to \infty,$$

then there is $X \in L^2(\Omega, \mathscr{F}, \mathbf{P})$ such that

$$||X_m - X|| \to 0 \textit{ as } m \to \infty.$$

Proof The assumption means that, given any $\varepsilon > 0$, we can find and index $N(\varepsilon)$ such that $||X_n - X_m|| \leq \varepsilon$ for all $n, m \geq N(\varepsilon)$. Since we can do this for any ε, we can do it for $\varepsilon = 2^{-1}, 2^{-2}, 2^{-3}, \ldots$. Thus, if $\varepsilon = 2^{-\ell}$, there is an index N_ℓ such that

$$||X_n - X_m|| \leq 2^{-\ell} \text{ for all } n, m \geq N_\ell. \tag{4.4}$$

Define $M_\ell = \max(N_1, N_2, \ldots, N_\ell)$. Then $||X_{M_{\ell+1}} - X_{M_\ell}|| \leq 2^{-\ell}$. But then

$$\mathbf{E}|X_{M_{\ell+1}} - X_{M_\ell}| \leq ||X_{M_{\ell+1}} - X_{M_\ell}|| \leq 2^{-\ell}.$$

Hence

$$\sum_{\ell=1}^{\infty} \mathbf{E}|X_{M_{\ell+1}} - X_{M_\ell}| \leq 1,$$

and, by Fubini's theorem,

$$\mathbf{E} \sum_{\ell=1}^{\infty} |X_{M_{\ell+1}} - X_{M_\ell}| < \infty.$$

Therefore, the random variable $sum_{\ell=1}^{\infty} |X_{M_{\ell+1}} - X_{M_\ell}|$ is finite with probability 1. Put it otherwise, there is a set $A \in \mathscr{F}$, with $\mathbf{P}(A) = 0$, such that

$$\sum_{\ell=1}^{\infty} |X_{M_{\ell+1}}(\omega) - X_{M_\ell}(\omega)| < \infty, \quad \omega \in A^c.$$

This means that

$$\lim_{n\to\infty}\sum_{\ell=1}^{n-1}(X_{M_{\ell+1}}(\omega)-X_{M_\ell}(\omega)) \text{ exists for all } \omega\in A^c.$$

But the sum is just $X_{M_n}(\omega)-X_{M_1}(\omega)$. We have thus proved that for all $\omega\in A^c$ there is $X(\omega)$ such that $X_{M_n}(\omega)\to X(\omega)$ as $n\to\infty$. Since this is a limit, the function X is a random variable. Fix ℓ and use (4.4) again with $m\geq M_\ell$, $n=M_r$, for $r\geq\ell$:

$$\mathbf{E}|X_m-X_{M_r}|^2\leq 4^{-\ell},\quad m\geq M_\ell,\quad r\geq\ell.$$

By Fatou's lemma, for all $m\geq M_\ell$,

$$\mathbf{E}|X_m-X|^2=\mathbf{E}\varliminf_{r\to\infty}|X_m-X_{M_r}|^2\leq\varliminf_{r\to\infty}\mathbf{E}|X_m-X_{M_r}|^2\leq 2^{-\ell}.$$

Since $||X_m-X||\geq||X||-||X_m||$, we have $||X||\leq||X_m||+2^\ell<\infty$, hence $X\in L^2(\Omega,\mathscr{F},\mathbf{P})$. Also, $\lim_{m\to\infty}||X_m-X||=0$, as required. □

4.7 Conditioning

4.7.1 Naïve conditioning

Naïve conditioning was defined in $\boxed{2.1}$. Namely, if $(\Omega,\mathscr{F},\mathbf{P})$ is a probability space and $B\in\mathscr{F}$, $\mathbf{P}(B)\neq 0$, we let

$$\mathbf{P}(A|B)=\mathbf{P}(AB)/\mathbf{P}(B),\quad A\in\mathscr{F},$$

be a new probability, called $\mathbf{P}$ conditional on B. We easily see that $(\Omega,\mathscr{F},\mathbf{P}(\cdot|B))$ is a new probability space.

The problem is that, in many cases, we want to condition with respect to an event B that has probability zero. This can be done. Loosely speaking, what saves us is the fact that if $\mathbf{P}(B)=0$ then $\mathbf{P}(AB)=0$ as well.

We can use many methods for defining conditioning in a more general sense. We adopt a geometric approach. In Chapter 8 you will see another approach.

4.7.2 Geometry

Consider the Euclidean space $\mathbb{R}^d$. Let $\mathbb{G}$ be a linear subspace of it. Let $||x-y||$ stand for the Euclidean distance between $x,y\in\mathbb{R}^d$, i.e.

$$||x-y||:=\sqrt{(x_1-y_1)^2+\cdots+(x_d-y_d)^2},$$

and

$$\langle x,y\rangle:=x_1y_1+\cdots+x_dy_d.$$

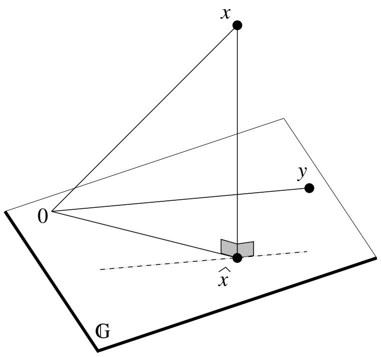
Thus $\langle x,y\rangle/||x||\,||y||$ is the cosine of the angle between x and y. We wish to project an $x\in\mathbb{R}^d$ onto $\mathbb{G}$:

Lemma 4.7. *Given any $x\in\mathbb{R}^d$ there is a unique $\widehat{x}\in\mathbb{G}$ such that*

$$||x-\widehat{x}||=\min_{y\in\mathbb{G}}||x-y||.$$

If we think geometrically, the answer is obvious: this $\widehat{x}$ is such that $x-\widehat{x}$ is orthogonal to $\mathbb{G}$; in other words, the angle between $x-\widehat{x}$ and any $y\in\mathbb{G}$ is a right angle, i.e. its cosine is zero:

$$\langle x-\widehat{x},y\rangle=0,\quad y\in\mathbb{G}.$$



Equivalently,

$$\langle \widehat{x}, y \rangle = \langle x, y \rangle, \quad y \in \mathbb{G}. \tag{4.5}$$

Since $\mathbb{G}$ is a linear space, (4.5) defines $\widehat{x}$ uniquely. To see this directly, assume that there are two elements $\widehat{x}, \widehat{x}' \in \mathbb{G}$ satisfying (4.5). Then $\langle \widehat{x}, y \rangle = \langle \widehat{x}', y \rangle$ for all $y \in \mathbb{G}$. Thus, $\langle \widehat{x} - \widehat{x}', y \rangle = 0$, for all $y \in \mathbb{G}$. In particular, for $y = \widehat{x} - \widehat{x}'$, we have $\langle \widehat{x} - \widehat{x}', \widehat{x} - \widehat{x}' \rangle = 0$, which immediately gives $\widehat{x} = \widehat{x}'$.

Proof of Lemma 4.7: Define (uniquely) $\widehat{x} \in \mathbb{G}$ by (4.5). Obviously,

$$||\widehat{x} - y||^2 = \langle \widehat{x} - y, \widehat{x} - y \rangle \geq 0, \quad y \in \mathbb{G}.$$

We have

$$\begin{aligned} 0 \leq \langle \widehat{x} - y, \widehat{x} - y \rangle &= ||\widehat{x}||^2 + ||y||^2 - 2\langle \widehat{x}, y \rangle \\ &= ||\widehat{x}||^2 + ||y||^2 - \langle x, y \rangle - \langle \widehat{x}, y \rangle, \quad y \in \mathbb{G}, \end{aligned}$$

where we used (4.5). We now use two things. First,

$$||x - y||^2 = ||x||^2 + ||y||^2 + 2\langle x, y \rangle,$$

and second,

$$||x||^2 = ||(x - \widehat{x}) + \widehat{x}||^2 = ||x - \widehat{x}||^2 + ||\widehat{x}||^2 + 2\langle x - \widehat{x}, x \rangle = ||x - \widehat{x}||^2 + ||\widehat{x}||^2,$$

because $\langle x - \widehat{x}, x \rangle = 0$, by (4.5). Combining the last two displays our inequality becomes

$$0 \leq ||x - y||^2 - ||x - \widehat{x}||^2, \quad y \in \mathbb{G},$$

which is what we want. □

Notation: We will write

$$\widehat{x} = \text{proj}(x|\mathbb{G}),$$

to remind ourselves that $\widehat{x}$ is obtained by projecting x onto $\mathscr{G}$.

EXERCISE 33. Let $\mathbb{H}$ be a linear subspace of $\mathbb{G}$. Then

$$\text{proj}\left(\text{proj}(x \mid \mathbb{G}) \mid \mathbb{H}\right) = \text{proj}(x \mid \mathbb{H}).$$

(This should be geometrically ‘obvious’: it is an exercise in right angles.)

4.7.3 Conditional probability

Suppose now that $\mathscr{G} \subset \mathscr{F}$ is a σ-field. Then $L^2(\Omega, \mathscr{G}, \mathbf{P})$ is a linear subspace of $L^2(\Omega, \mathscr{F}, \mathbf{P})$. By analogy to the geometric picture above, given $X \in L^2(\Omega, \mathscr{F}, \mathbf{P})$, we wish to define $\widehat{X} \in L^2(\Omega, \mathscr{G}, \mathbf{P})$ such that

$$||X - \widehat{X}|| = \min_{Y \in L^2(\Omega,\mathscr{G},\mathbf{P})} ||X - Y||.$$

Owing to the structure of $L^2(\Omega, \mathscr{G}, \mathbf{P})$, this can be done:

Theorem 4.3. *Let $\mathscr{G} \subset \mathscr{F}$ be a σ-field. Given $X \in L^2(\Omega, \mathscr{F}, \mathbf{P})$, there exists $\widehat{X} \in L^2(\Omega, \mathscr{G}, \mathbf{P})$ such that one of the following two equivalent statements holds:*

(i)

$$||X - \widehat{X}|| = \inf_{Y \in L^2(\Omega,\mathscr{G},\mathbf{P})} ||X - Y||. \tag{4.6}$$

(ii)

$$\mathbf{E}XY = \mathbf{E}\widehat{X}Y, \quad \textit{for all } Y \in L^2(\Omega, \mathscr{G}, \mathbf{P}). \tag{4.7}$$

Furthermore, $\widehat{X}$ is almost surely unique, in the sense that if $\widehat{X}' \in L^2(\Omega, \mathscr{G}, \mathbf{P})$ satisfies (i) or (ii) then $\mathbf{P}(\widehat{X} = \widehat{X}') = 1$.

Proof [sketch]: We only need to prove existence of such an $\widehat{X}$. The rest of the proof is as in the previous section. Existence is guaranteed by Lemma 4.4. Indeed, by the definition of inf, we can choose $Y_n \in L^2(\Omega, \mathscr{G}, \mathbf{P})$ such that $||X - Y_n||$ converges to the infimum in (4.6). It is easy to see that $\sup_{m,n \geq N} ||Y_n - Y_m|| \to 0$ as $N \to \infty$, therefore, by Lemma 4.4, we have an $\widehat{X} \in L^2(\Omega, \mathscr{G}, \mathbf{P})$ such that $||Y_n - \widehat{X}|| \to 0$ as $n \to \infty$. This $\widehat{X}$ does the job. □

We use the following notation:

$$\widehat{X} = \mathbf{E}_{\mathbf{P}}(X|\mathscr{G}).$$

The notation reminds us that $\widehat{X}$ depends both on $\mathscr{G}$ and on $\mathbf{P}$. If $\mathbf{P}$ is implicitly understood then we just write $\widehat{X} = \mathbf{E}(X|\mathscr{G})$.

The next exercise justifies using the symbol $\mathbf{E}$ for this projection. Indeed, if the σ-field $\mathscr{G}$ is the smallest possible, then the projection is a constant, the expectation of X:

EXERCISE 34. Show that $\mathbf{E}(X|\{\varnothing, \Omega\}) = \mathbf{E}X$. (Hint: $\mathbf{E}(X - t)^2$ is minimised by $t = \mathbf{E}X$.)

What happens if X is $\mathscr{G}$ measurable?

EXERCISE 35. If $X \in L^2(\Omega, \mathscr{G}, \mathbf{P})$ then $\mathbf{E}(X|\mathscr{G}) = X$.

We can also show the double projection property:

EXERCISE 36. If $\mathscr{H} \subset \mathscr{G} \subset \mathscr{F}$ are all σ-fields then

$$\mathbf{E}(\mathbf{E}(X \mid \mathscr{G}) \mid \mathscr{H}) = \mathbf{E}(X \mid \mathscr{H}).$$

Suppose now that $A \in \mathscr{F}$. Then we define

$$\mathbf{P}(A \mid \mathscr{G}) := \mathbf{E}(\mathbf{1}_A \mid \mathscr{G}),$$

and call this conditional probability of the event A given the σ-field $\mathscr{G}$. By definition, for each $A \in \mathscr{F}$, the conditional probability $\mathbf{P}(A \mid \mathscr{G})$ is a random variable, specifically it is $\mathscr{G}$-measurable. Also, by linearity, if $A_1 A_2 = \varnothing$,

$$\mathbf{P}(A_1 \cup A_2|\mathscr{G}) = \mathbf{P}(A_1 \mid \mathscr{G}) + \mathbf{P}(A_2 \mid \mathscr{G}),$$

except on a set of probability zero. The question is: Is $\mathscr{F} \ni A \mapsto \mathbf{P}(A \mid \mathscr{G})$ a probability for almost every $\omega \in \Omega$? The answer is not obvious because $\mathbf{P}(A \mid \mathscr{G})$ is defined uniquely up to an event N_A of probability zero. Since there are uncountably many A's in $\mathscr{F}$, it is not clear we can simultaneously define all the random variables $\mathbf{P}(A \mid \mathscr{G})$ up to an event of probability zero. In special cases this can be done.

4.7.4 Conditional law

Here is one such special case: Let X, Y be random variables on $(\Omega, \mathscr{F}, \mathbf{P})$ with finite second moments. We define

$$\mathbf{P}(X \in B \mid Y) := \mathbf{P}(X^{-1}(B) \mid \sigma(Y)), \quad B \in \mathscr{B}.$$

While we can certainly do this rigorously for each fixed $B \in \mathscr{B}$, it is not immediate that for each $\omega \in \Omega$, this is a probability in B. It can be shown that it actually is, in the sense that we can choose $\mathbf{P}(X \in B \mid Y)$ so that it is a probability as a function of B and simultaneously a measurable function of Y. We call this object conditional law of X given Y.

The conditional distribution function of X given Y is defined as

$$\mathbf{P}(X \leq x \mid Y).$$

To be honest with the definitions, we must make sure they reduce to the usual naïve ones from elementary probability. We consider two cases.

Case 1: (X, Y) is absolutely continuous

Lemma 4.8. *Suppose that (X, Y) is absolutely continuous with joint density $f(x, y)$. Let $f_2(y) = \int_{\mathbb{R}} f(x, y)dx$ be the density of Y. Define the conditional density of X given Y by*

$$f_{X|Y}(x|y) := \begin{cases} \dfrac{f(x,y)}{f_2(y)}, & \text{if } f_2(y) \neq 0, \\ 0, & \text{otherwise.} \end{cases}$$

For $B \in \mathscr{B}$, let

$$g_B(y) := \int_B f_{X|Y}(x|y)dx.$$

Then

$$P(X \in B|Y) = g_B(Y),$$

for all ω except on a set of probability zero.

Proof We need to show that (see (4.7))

$$\mathbf{E}\mathbf{1}(X \in B)W = \mathbf{E}g_B(Y)W,$$

for all $W \in L^2(\Omega, \sigma(Y), \mathbf{P})$. It is enough to show this for al, W of the form $W = \mathbf{1}(Y \in C)$, $C \in \mathscr{B}$:

$$\mathbf{E}\mathbf{1}(X \in B)\mathbf{1}(Y \in C) = \mathbf{E}g_B(Y)\mathbf{1}(Y \in C).$$

But the left hand side equals

$$\int_{\mathbb{R}^2} \mathbf{1}_B(x)\mathbf{1}_C(y)f_{X,Y}dxdy.$$

The right hand side equals

$$\int_R g_B(y)\mathbf{1}_C(y)dy.$$

These two quantities are the same, by Fubini's theorem. □

Case 2: (X, Y) is discrete

Suppose that (X, Y) takes values in a discrete set $S_1 \times S_2$. As usual, define its probability mass function and marginal mass functions by

$$p(x, y) := \mathbf{P}((X, Y) = (x, y)), \quad p_1(x) := \mathbf{P}(X = x), \quad p_2(y) := \mathbf{P}(Y = y), \quad (x, y) \in S_1 \times S_2.$$

We then have a simple formula for the conditional law of X given Y, precisely the one we expect:

Lemma 4.9. *Let (X, Y) be a random variable on $(\Omega, \mathscr{F}, \mathbf{P})$ with values in the discrete set $(S_1 \times S_2, 2^{S_1 \times S_2})$. Define the conditional probability mass function of X given Y by*

$$p_{X|Y}(x|y) := \begin{cases} \dfrac{p(x,y)}{p_2(y)}, & \textit{if } p_2(y) \neq 0, \\ 0, & \textit{otherwise.} \end{cases}$$

Let $\mathbf{P}(X = x|Y)$ be defined as the projection of $\mathbf{1}(X = x)$ onto $L^2(\Omega, \sigma(Y), \mathbf{P})$. Then[a]

$$\mathbf{P}(X = x|Y) = p_{X|Y}(x|Y).$$

Proof We need to verify that $p_{X|Y}(x|Y)$ satisfies (4.7), namely,

$$\mathbf{E}\mathbf{1}(X = x)W = \mathbf{E}p_{X|Y}(x|Y)W, \quad W \in L^2(\Omega, \sigma(Y), \mathbf{P}).$$

We work with the case where S_2 is a finite set, the more general case being similar. We can then write Y as a finite sum of the form

$$Y = \sum_y y\mathbf{1}(Y = y)$$

where the A_k are mutually disjoint with $\mathbf{P}(A_k) > 0$. Any W which is $\sigma(Y)$-measurable is a function of Y, and so we can take

$$W = \sum_y g(y)\mathbf{1}(Y = y).$$

We have

$$\begin{aligned} \mathbf{E}[p_{X|Y}(x|Y)W] &= \sum_y g(y)\mathbf{E}[p_{X|Y}(x|Y)\mathbf{1}(Y = y)] \\ &= \sum_y g(y)\mathbf{E}[p_{X|Y}(x|y)\mathbf{1}(Y = y)] = \sum_y g(y)p_{X|Y}(x|y)\mathbf{P}(Y = y). \end{aligned}$$

If $\mathbf{P}(Y = y) = 0$ then $p_{X|Y}(x|y) = 0$. Otherwise, $p_{X|Y}(x|y) = p(x, y)/p_2(y)$, and so

$$\mathbf{E}[p_{X|Y}(x|Y)W] = \sum_y g(y)p(x, y) = \mathbf{E}XW,$$

as required. □

[a] The notation $p_{X|Y}$ is terrible. We only use it out of some respect to the undergraduate probability courses. The reason that the notation is terrible is that in the subsript '$X|Y$' in $p_{X|Y}(x|Y)$ plays a merely cosmetic rôle, as opposed to the essential rôle played by the last variable Y inside the parenthesis.

4.8 Gaussian variables

4.8.1 The Gaussian law

The motivation of the Gaussian probability comes from the central limit theorem (which, long time ago, was known as the "law of errors"). This was stated, without proof, in $\boxed{2.3.2}$. We work heuristically in order to motivate the definitions. Let $S_n = \xi_1 + \cdots + \xi_n$ be the sum of n independent indicator (also known as Bernoulli) random variables ξ_i with $\mathbf{P}(\xi_i = 1) = p$ for all i. Let $\widehat{S}_n = S_n - \mathbf{E}S_n$. Then, as $n \to \infty$, the distribution function of $\widehat{S}_n/\sqrt{n}$ converges to an absolutely continuous distribution function with a famous formula. Let X be a random variable with such a distribution function. Note that

$$\frac{\widehat{S}_{2n}}{\sqrt{n}} = \frac{1}{\sqrt{2}}\frac{\widehat{S}_n}{\sqrt{n}} + \frac{1}{\sqrt{2}}\frac{\widehat{S}'_n}{\sqrt{n}},$$

where $\widehat{S}'_n = \xi_{n+1} + \cdots + \xi_{2n}$ is a random variable with the same law as $\widehat{S}_n$. So the distribution of $\frac{\widehat{S}'_n}{\sqrt{n}}$ will also converge to the distribution of X. Moreover, $\widehat{S}'_n$, $\widehat{S}_n$ are independent. Therefore, if X_1, X_2 are independent random variables with the same law as X, then we expect that

$$X = \frac{X_1 + X_2}{\sqrt{2}}.$$

Even if we do not know what this famous distribution is, it should be such that this "addition law" holds. From this, we can discover its density. One way to do that is by imposing the extra assumption (which is not necessary) that the generating function of X exists for all θ: $M(\theta) = \mathbf{E}e^{\theta X}$. Then

$$M(\theta) = M(\theta/\sqrt{2})^2.$$

Letting $\theta = \sqrt{\eta}$ and taking logarithms, we have

$$\log M(\sqrt{\eta}) = 2 \log M(\sqrt{\eta/2}).$$

So, if we temporarily let $m(\eta) = \log M(\sqrt{(}\eta))$ we have

$$m(\eta) = 2m(\eta/2).$$

So $m(0) = 0$ and with some work, we can actually find that the only continuous function satisfying the latter is linear: $m(\eta) = c\eta$. (This should be geometrically obvious.) Hence

$$M(\theta) = e^{c\theta^2}.$$

From Lemma 3.32 we know that the moments of X are given by the derivatives of M at 0. We find

$$M'(0) = 0, \quad M''(0) = 2c.$$

So $\mathbf{E}X = 0$ (as it should), while $\mathbf{E}X^2 = 2c$. So $c > 0$. Since $\mathbf{E}X = 0$, the second moment is the variance and we customarily denote it by σ^2. We arrive at

$$M(\theta) = e^{\frac{1}{2}\sigma^2\theta^2}.$$

We know that there is only one probability distribution with a given moment generating function. Instead of trying to figure out which one it is, let us reveal the result and then just verify that it is correct. We claim that the probability distribution corresponding to the last M is absolutely continuous with density

$$f(x) = Ce^{-x^2/2\sigma^2}.$$

Here, C is such that $\int_{\mathbb{R}} f(x)dx = 1$. This is the famous normal or Gaussian density with mean 0 and variance σ^2. It is called standard if $\sigma^2 = 1$.

EXERCISE 37. Show that $\int_{-\infty}^{\infty} e^{\theta x} f(x) dx = e^{\frac{1}{2}\sigma^2\theta^2}$. (Hint: Complete the square and use the definition of C.)

Now, to find C is a very important thing. It is based on the following:

Lemma 4.10.

$$\int_{\mathbb{R}} e^{-x^2} dx = \sqrt{\pi}.$$

(Liouville said that a Mathematician is someone for whom this integral is obvious.)

EXERCISE 38. Use Fubini's theorem to write $\left(\int_{\mathbb{R}} e^{-x^2} dx\right)^2 = \int_{\mathbb{R}^2} e^{-x^2-y^2} dxdy$ and do the latter integral using polar coördinates.

Using this we find that

$$C = \frac{1}{\sqrt{2\pi\sigma^2}}.$$

Therefore, the standard normal density is

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

We write $\mathcal{N}(0, 1)$ for the law of a random variable X with standard normal density. We write $\mathcal{N}(\mu, \sigma^2)$ for the law of $\sigma X + \mu$.

EXERCISE 39. Show that a density for $\mathcal{N}(\mu, \sigma^2)$ is

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.$$

EXERCISE 40. Show that if $X_i, i = 1, \ldots, d$ are independent and X_i having law $\mathcal{N}(\mu_i, \sigma_i^2)$ then $\sum_{i=1}^d X_i$ has law $\mathcal{N}(\sum_i \mu_i, \sum_i \sigma_i^2)$. Therefore linear combinations of independent normal variables are normal.

4.8.2 The multidimensional Gaussian random variable

We now pass on to defining a Gaussian (or normal) variable in $\mathbb{R}^d$ (i.e. a random vector).

> We say that $(X_1, \ldots, X_d)$ is Gaussian in $\mathbb{R}^d$ if, for all $a_1, \ldots, a_d \in \mathbb{R}$, the random variable $a_1 X_1 + \cdots + a_d X_d$ is normal.

The next lemma shows what the moment generating function of a normal vector is:

Lemma 4.11. *If* $(X_1, \ldots, X_d)$ *is Gaussian vector with*

$$\mu_j = \mathbf{E} X_j, \quad r_{jk} = \mathrm{cov}(X_j, X_k),$$

then

$$\mathbf{E} \exp \sum_{j=1}^d \theta_j X_j = \exp\left\{\sum_{j=1}^d \mu_j \theta_j + \frac{1}{2} \sum_{j=1}^d \sum_{k=1}^d r_{jk} \theta_j \theta_k\right\}.$$

Proof By definition $\sum_{j=1}^{d} \theta_j X_j$ should be normal, i.e. have a law $\mathcal{N}(\mu, \sigma^2)$, for some μ, σ^2. We have

$$\mu = \mathbf{E} \sum_{j=1}^{d} \theta_j X_j = \sum_{j=1}^{d} \mu_j \theta_j, \quad \sigma^2 = \operatorname{cov} \sum_{j=1}^{d} \theta_j X_j = \sum_{j=1}^{d} \sum_{k=1}^{d} \theta_j \theta_k \operatorname{cov}(X_j, X_k).$$

□

Since the moment generating function of a Gaussian vector is the exponential of a quadratic form, there is no better way to express it other than using Linear Algebra. To this end, we think of the elements x of $\mathbb{R}^d$ as column vectors. We use x' to denote transposition, i.e. the corresponding row when x is a column. And, of course, $(x')' = x$. Consider the mean (column) vector

$$\mu = (\mu_1, \ldots, \mu_d)'$$

and the symmetric covariance matrix

$$R = [r_{jk}].$$

Write also X for the column with entries $X_1, \ldots, X_d$. We then have

$$\mathbf{E} e^{\theta' X} = \exp\left\{ \theta' \mu + \frac{1}{2} \theta' R \theta \right\}.$$

This uniquely defines the law of $(X_1, \ldots, X_d)$. This law is denoted by $\mathcal{N}(\mu, R)$, where μ is the mean vector and R the covariance matrix.

Lemma 4.12. *A Gaussian vector $(X_1, \ldots, X_d)$ is absolutely continuous if and only if R is invertible. In this case, its density is given by*

$$f(x) = \frac{1}{\sqrt{(2\pi)^d \det(R)}} \exp\left(-\frac{1}{2} (x - \mu)' R^{-1} (x - \mu) \right).$$

Proof Assume $\mu = 0$, to ease notation. Assume R is invertible. It is easily seen that

$$R = \mathbf{E}(XX')$$

where XX' is a $d \times d$ matrix and $\mathbf{E}(XX')$ is the matrix formed by taking the expectations of the entries of XX'. This R has two important properties:
(i) it is symmetric (obviously);
(ii) it is positive semi-definite, i.e. the quadratic form

$$u' R u = \sum_k \sum_\ell R_{k\ell} u_k u_\ell \geq 0$$

for all values of the variables, positive or negative. The reason for the latter is that $u'Ru$ is the expectation of a non-negative quantity:

$$u' R u = u' \mathbf{E} X X' u = \mathbf{E}(X'u)'(X'u) = \mathbf{E}(X'u)^2 \geq 0.$$

Standard linear algebra shows that R has exactly d (counting multiplicities) non-negative eigenvalues. (In fact, they are strictly positive, due to the invertibility of R which is tantamount to $\det(R) \neq 0$). Furthermore, the eigenvectors can be chosen to be orthonormal. Letting U be the matrix whose columns are these d orthonormal eigenvectors and Λ the diagonal matrix with the eigenvalues in its diagonal, we have

$$RU = U\Lambda,$$

from the very definition of the eigenvectors. Now

$$U' = U^{-1},$$

hence

$$R = U\Lambda U' = U\Lambda^{1/2}\Lambda^{1/2}U' = PP',$$

with

$$P = U\Lambda^{1/2}.$$

The matrix P is non-singular and is called the square root of R. We now define new random variables Z by

$$X = PZ\,.$$

The thing to observe is that the covariance matrix of Z is

$$\mathrm{cov}(Z) = \mathbf{E}ZZ' = \mathbf{E}P^{-1}XX'P'^{-1} = P^{-1}R'P'^{-1} = P^{-1}PP'P'^{-1} = I,$$

I being the identity matrix. Thus

$$\mathbf{E}e^{\theta' Z} = e^{\theta'\theta/2} = \prod_{j=1}^{d} e^{\theta_j^2/2} = \prod_{j=1}^{d} \mathbf{E}e^{\theta_j Z_j},$$

implying that the components of Z are independent standard normal. Hence the density g of Z is product:

$$g(z) = \prod_j \frac{1}{\sqrt{2\pi}} e^{-z_j^2/2}.$$

Now $X = PZ$, hence its density f is computed easily by

$$f(x) = g(P^{-1}x)/|\det(P)|,$$

which yields the desired formula. □

If R has determinant zero then it is possible to "reduce the dimension" of the random vector X so that the density exists. In fact,

Lemma 4.13. *The support of X is the range of its covariance matrix R.*

Proof Suppose that R has rank r. Then it has $d-r$ eigenvalues at zero, so that the matrix Λ consists of a $d-r$ size block of zeros and the remaining non-zero eigenvalues. Hence now $R = PP'$, where P is a $d \times r$ matrix with rank r. We try again to find Z so that

$$X = PZ,$$

where Z is an r-dimensional random vector with density. If we manage to do this we will finish, since the range of P is the range of R. Observing that $P'P$ is an $r \times r$ non-singular matrix, we pre-multiply by it to get $P'PZ = P'X$, hence $Z = (P'P)^{-1}P'X$. So if we define Z this way, we see that

$$\mathbf{E}ZZ' = (P'P)^{-1}P'RP(P'P)^{-1} = (P'P)^{-1}P'PP'P(P'P)^{-1} = I,$$

i.e. Z is a collection of r independent standard normal variables. The formula for Z is the formula that solves a full-rank over-estimated linear system. We usually write $Z = P^{+}X$ and call P^{+} the pseudo-inverse of P. It remains to show that every linear function of X is a linear function of Z. Let

$$F = \lambda' X a \quad , G = \lambda' P Z$$

with $Z = P^+ X$. Consider

$$F - G = \lambda'(X - PZ).$$

Then

$$\mathbf{E}(F-G)^2 = \mathbf{E}\lambda'(X - PZ)(X' - Z'P')\lambda = \mathbf{E}\lambda'(XX' - XZ'P' - PZX' + PZZ'P')\lambda.$$

But

$$\mathbf{E}ZX' = (P'P)^{-1}P'\mathbf{E}XX' = (P'P)^{-1}P'PP' = P'.$$

So

$$\mathbf{E}(F-G)^2 = \lambda'(PP' - PP' - PP' + PP')\lambda = 0.$$

□

Terminology: If X has law $\mathcal{N}(0, R)$ with R having rank r then $\sum_j X_j^2$ is called χ^2 with r degrees of freedom.

4.8.3 Conditional Gaussian law

If it appears that we've done a lot of Linear Algebra, then this is because it is so: Dealing with Gaussian random variables (and processes!) is mostly dealing with Linear Algebra (or Linear Analysis!).

Without proof, we mention the following:

Lemma 4.14. *Let $(X; Y_1, \ldots, Y_d)$ be a Gaussian random variable in $\mathbb{R}^{1+d}$. Then the conditional law $\mathbf{P}(X \in \cdot | Y_1, \ldots, Y_d)$ is normal with mean $\mathbf{E}(X|Y_1, \ldots, Y_d)$ and deterministic covariance matrix. Moreover, $\mathbf{E}(X|Y_1, \ldots, Y_d)$ is a linear function of $Y_1, \ldots, Y_d$).*

EXERCISE 41. Let X, Y be independent standard Gaussian variables. Based on the last part of the lemma above, compute $\mathbf{E}(X + 2Y|3X - 4Y)$. (Hint: Make sure that (4.7) is satisfied.)

SMSTC (2007/08)

Probability

5: Important special distributions

Takis Konstantopoulos[a]

www.smstc.ac.uk

Contents

5.1 Outline

We discuss several special random variables, and simple relationships between their distributions.

5.2 Binomial, Poisson, and multinomial

Consider the coin tossing experiment, i.e. consider a sequence $\xi_1, \xi_2, \dots$ of i.i.d. random variables with

$$\mathbf{P}(\xi_1 = 1) = p, \quad \mathbf{P}(\xi_1 = 0) = 1 - p.$$

The law of

$$S_n = \xi_1 + \cdots + \xi_n$$

is called Binomial with parameters n and p. From this we have

$$\mathbf{E}S_n = np, \quad \operatorname{var} S_n = n \operatorname{var} \xi_1 = np(1-p),$$

and, since $\xi_{n+1} + \xi_{n+m}$ is Binomial with parameters m and p, and independent of S_n, we have that the sum of two independent Binomial random variables with the same p is again Binomial. We also have

$$\mathbf{P}(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \dots, n,$$

and, from the Binomial theorem,

$$\mathbf{E}e^{\theta S_n} = (pe^\theta + 1 - p)^n.$$

EXERCISE 42. Compute the moment generating function of $n^{-1/2}(S_n - \mathbf{E}S_n)$ and show that, as $n \to \infty$, it converges to the moment generating function of a Gaussian random variable.

EXERCISE 43. Use Chernoff's inequality to estimate $\mathbf{P}(S_n > n(p + x))$ for $x > 0$.

By letting p vary with n and taking limits we obtain a different fundamental law, the Poisson law. Specifically,

Lemma 5.1. *If $p_n = \frac{\lambda}{n} + o(1/n)$, as $n \to \infty$, then*

$$\mathbf{P}(S_n = k) \to \frac{\lambda^k}{k!} e^{-\lambda},$$

for all $k = 0, 1, 2, \ldots$.

Proof Use Stirling's formula. □

This is the Poisson law with parameter λ. The Poisson law is fundamental when, roughly speaking, we deal with independent rare events.

Let X be a Poisson random variable. Then We have

$$\mathbf{E}e^{\theta X} = \sum_{k=0}^{\infty} \frac{(\lambda e^{\theta})^k}{k!} e^{-\lambda} = e^{\lambda(e^{\theta}-1)}.$$

Differentiating a couple of times, we find

$$\mathbf{E}X = \lambda, \quad \operatorname{var} X = \lambda.$$

Furthermore,

Lemma 5.2. *If $X_1, X_2, \ldots$ are independent Poisson random variables with parameters $\lambda_1, \lambda_2, \ldots$ such that $\sum \lambda_k < \infty$ then $\sum_k X_k$ is Poisson with parameter $\sum_k \lambda_k$.*

Proof We prove this for a finite number n of random variables (which is enough by the way we construct a probability on an infinite product). The moment generating function of $X_1 + \ldots + X_n$ is

$$\prod_{k=1}^{n} e^{\lambda_k(e^{\theta}-1)} = e^{(e^{\theta}-1)\sum_{k=1}^{n}\lambda_k}$$

and this is the moment generating function of a Poisson law with parameter $\sum_k \lambda_k$. □

Next we consider conditional probabilities:

Lemma 5.3. *Suppose that X_1, X_2 are independent Poisson with parameters λ_1, λ_2. Then, conditional on $X_1 + X_2 = n$, we have that X_1 is Binomial with parameters n, $\lambda_1/(\lambda_1 + \lambda_2)$.*

Proof Elementary conditioning: For $0 \le k \le n$,

$$\begin{aligned}\mathbf{P}(X_1 = k | X_1 + X_2 = n) &= \frac{\mathbf{P}(X_1 = k, X_2 = n-k)}{\mathbf{P}(X_1 + X_2 = n)} \\ &= \frac{\frac{\lambda_1^k}{k!} e^{-\lambda_1} \, \frac{\lambda_2^{n-k}}{(n-k)!} e^{-\lambda_2}}{\frac{(\lambda_1+\lambda_2)^n}{n!} e^{-\lambda_1-\lambda_2}} \\ &= \binom{n}{k} \left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^k \left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{n-k}.\end{aligned}$$

□

Generalising this, we have:

Lemma 5.4. *Suppose that $X_1, X_2, \dots, X_d$ are independent Poisson with parameters $\lambda_1, \lambda_2, \dots, \lambda_d$, respectively. Then, conditionally on $\sum_k X_k = n$, the random vector $(X_1, \dots, X_d)$ has law given by*

$$\mathbf{P}(X_1 = n_1, \dots, X_d = n_d \mid \sum_k X_k = n) = \binom{n}{n_1, \dots, n_d} \left(\frac{\lambda_1}{\lambda}\right)^{n_1} \cdots \left(\frac{\lambda_d}{\lambda}\right)^{n_d},$$

where $(n_1, \dots, n_d)$ are nonnegative integers with sum equal to n, and $\lambda = \sum_k \lambda_k$, and where

$$\binom{n}{n_1, \dots, n_d} = \frac{n!}{n_1! \cdots n_d!}$$

EXERCISE 44. Show this.

The symbol $\binom{n}{n_1, \dots, n_d}$ is the multinomial coefficient since it appears in the algebraic identity known as multinomial theorem:

$$(x_1 + \cdots + x_d)^n = \sum \binom{n}{n_1, \dots, n_d} x_1^{n_1} \cdots x_d^{n_d}.$$

The sum is taken over all nonnegative integers $(n_1, \dots, n_d)$, with sum equal to n.

The random variable $(X_1, \dots, X_d)$ with values in $\mathbb{Z}_+^d$ is said to have a multinomial distribution with parameters $d, n, p_1, \dots, p_d$ (where $p_1 + \cdots + p_d = 1$, so one of them is superfluous) if

$$\mathbf{P}(X_1 = n_1, \dots, X_d = n_d) = \binom{n}{n_1, \dots, n_d} p_1^{n_1} \cdots p_d^{n_d},$$

where $(n_1, \dots, n_d)$ are nonnegative integers with sum equal to n,

Of course, a multinomial distribution with parameters $2, n, p, 1-p$ is a binomial distribution with parameters n, p.

5.3 Thinning

Suppose an urn contains n balls. There are d colours available. Let the colours be denoted by $c_1, c_2, \dots, c_d$. To each ball assign colour c_i with probability p_i, independently from ball to ball. Let S_n^i be the number of balls that have colour c_i, $1 \le i \le d$. It is easy to see that $(S_n^1, \dots, S_n^d)$ has a multinomial law:

$$\mathbf{P}(S_n^1 = k_1, \dots, S_n^d = k_d) = \binom{n}{k_1, \dots, k_d} p_1^{k_1} \cdots p_d^{k_d},$$

where $(k_1, \dots, k_d)$ are nonnegative integers summing up to n. Clearly,

$$S_n^1 + \cdots S_n^d = n,$$

so the random variable $S_n^1, \dots, S_n^d$ cannot be independent. The question is:

> Suppose that the number of balls is itself a random variable, independent of everything else. Is there a way to choose the law of this random variable so that the above variables are independent?

To put the problem in precise mathematical terms, let $\xi_1, \xi_2, \dots$ be i.i.d. random colours, i.e. random variables taking values in $\{c_1, \dots, c_d\}$ such that

$$\mathbf{P}(\xi_1 = c_i) = p_i, \quad 1 \le i \le d.$$

Let

$$S_n^i := \sum_{k=1}^n \mathbf{1}(\xi_k = c_i).$$

(In physical terms, S_n^i denotes precisely what we talked about earlier using a more flowery language.) Now, independent of the sequence $\xi_1, \xi_2, \ldots$, let N be a random variable with values in $\mathbb{Z}_+$. The problem is to find its law so that

$$S_N^1, \ldots, S_N^d \text{ are independent random variables.} \qquad (\star)$$

It turns out that this is a characterising property of the Poisson law. We will contend ourselves by proving one direction:

Lemma 5.5. *If N is Poisson then $(\star)$ holds. Moreover, if N has expectation λ, then S_N^i is also Poisson with expectation λp_i.*

EXERCISE 45. Prove the last lemma.

5.4 Geometric

A random variable X with values in $\mathbb{N}$ is geometric if it has the memoryless property:

> For each $k \in \mathbb{N}$, the conditional distribution of $X - k$ given $\{X > k\}$ is the same as the distribution of X:
>
> $$\mathbf{P}(X - k = n | X > k) = \mathbf{P}(X = n).$$

Think of X as a random time, e.g. the day on which a certain volcano will erupt. The property above says that if by day k the volcano has not erupted then the remaining time $X - k$ has the same law as X, no matter how large k is.

Lemma 5.6. *If $q = \mathbf{P}(X > 1)$ then*

$$\mathbf{P}(X > k) = q^k, \quad k \in \mathbb{N}$$

and

$$\mathbf{P}(X = k) = pq^{k-1},$$

where $p = 1 - q$.

Proof By the property of X,

$$\mathbf{P}(X > k + n | X > k) = \mathbf{P}(X > n),$$

for all k, n, which means that

$$\mathbf{P}(X > k + 1) = \mathbf{P}(X > k)\mathbf{P}(X > 1).$$

Iterating this, we find

$$\mathbf{P}(X > k) = \mathbf{P}(X > 1)^k, \quad k = 0, 1, \ldots.$$

□

People refer to X as geometric with parameter p. The terminology is not standard because other people refer to $X - 1$ (which also has the memoryless property but takes values in $\mathbb{Z}_+$) as geometric with parameter q. A matter of taste, really.

It is easy to see that

Lemma 5.7. *If X is geometric in $\mathbb{N}$ with $\mathbf{P}(X = 1) = p$ then*

$$\mathbf{E}X = 1/p, \quad \operatorname{var} X = (1-p)/p^2, \quad \mathbf{E}e^{\theta X} = \frac{pe^\theta}{1-(1-p)e^\theta}.$$

EXERCISE 46. Do all that.

A concrete way to get a geometric random variable is by considering $\xi_1, \xi_2, \dots$ to be i.i.d. with

$$\mathbf{P}(\xi_1 = 1) = 1 - \mathbf{P}(\xi_1 = 0) = p$$

and by letting

$$X = \inf\{k \geq 1 : \ \xi_k = 1\}.$$

We have $\mathbf{P}(X < \infty) = 1$, so

$$X = \min\{k \geq 1 : \ \xi_k = 1\}$$

and

$$\mathbf{P}(X > k) = \mathbf{P}(\xi_1 = \cdots = \xi_k = 0) = (1-p)^k,$$

as required.

We have that

Lemma 5.8. *If $X_1, X_2, \dots, X_d$ are independent and geometric then $X = \min(X_1, \dots, X_d)$ is geometric.*

Proof

$$\begin{aligned}\mathbf{P}(X > k) &= \mathbf{P}(X_1 > k, \dots, X_d > k)\\ &= \mathbf{P}(X_1 > k) \cdots \mathbf{P}(X_d > k)\\ &= q_1^k \cdots q_d^k = (q_1 \cdots q_d)^k.\end{aligned}$$

□

EXERCISE 47. Let X, Y be independent and geometric. Show that

$$\mathbf{P}(X - Y > n | X > Y) = \mathbf{P}(X > n),$$

for all n, and interpret the formula.

5.5 Uniform

We have already seen, in detail, how to construct a uniform random variable, from first principles. Recall that U is uniform in the interval $[0,1]$ if $\mathbf{P}(U \leq x) = x$, $0 \leq x \leq 1$. More generally,

> X is uniform in $[a,b]$ if, for all intervals I, the probability $\mathbf{P}(X \in I)$ is proportional to the length of I.

Of course, if X is uniform in $[a,b]$, then $cX + d$ is uniform in the interval with endpoints $ca + d$ and $cb + d$.

Recall that if F is a distribution function and U is uniform in $[0,1]$ then $F^{-1}(U)$ is a random variable with distribution function F.

Lemma 5.9. *Let $p_1, \dots, p_d$ be positive numbers adding up to 1. Split the interval $[0,1]$ into consecutive intervals $I_1, \dots, I_d$ of lengths $p_1, \dots, p_d$, respectively. Let $U_1, \dots, U_n$ be i.i.d. uniform in $[0,1]$. Let*

$$S_n^i = \sum_{k=1}^{n} \mathbf{1}(U_k \in I_i), \quad 1 \le i \le d.$$

Then $(S_n^1, \dots, S_n^d)$ has a multinomial law. In particular, S_n^i is Binomial with parameters n, p_i.

EXERCISE 48. Show this last lemma.

EXERCISE 49. Let $U_1, \dots, U_d$ be i.i.d. uniform in $[0,1]$. Compute the probability $\mathbf{P}(U_1 < U_2 < \cdots < U_d)$.

EXERCISE 50. Consider a stick of length 1 and break it into 3 pieces, by choosing the two break points at random. Find the probability that the 3 smaller sticks can be joined to form a triangle.

EXERCISE 51. Pick a random variable U_1 uniform in $[0,1]$. Let U_2 be the midpoint of the interval $[0, U_1]$ or of $[U_1, 1]$, with equal probability. Continue in the same manner and define U_3 to be the midpoint of $[0, U_2]$ or of $[U_2, 1]$, with equal probability. Show that the $x \mapsto \lim_{n\to\infty} \mathbf{P}(U_n \le x)$ is continuous but not absolutely so.

5.6 Exponential

A random variable T with values in $\mathbb{R}_+$ is exponential if it has the memoryless property:

For all $t, s > 0$,

$$\mathbf{P}(T - t > s | T > t) = \mathbf{P}(T > s).$$

Lemma 5.10. *If T is exponential then there is $\lambda > 0$ such that*

$$\mathbf{P}(T > t) = e^{-\lambda t}, \quad t \ge 0.$$

Proof Implicit in the definition is that $\mathbf{P}(T > t) > 0$ for all t. Hence $\alpha := \mathbf{P}(T > 1) \in (0,1)$. We have

$$\mathbf{P}(T > t + s) = \mathbf{P}(T > t)\mathbf{P}(T > s)$$

for all t, s. Using this and induction, we have that, for all $n \in \mathbb{N}$,

$$\mathbf{P}(T > nt) = \mathbf{P}(T > t)^n.$$

This gives that, for all $m \in \mathbb{N}$,

$$\mathbf{P}(T > 1 = m(1/m)) = \mathbf{P}(T > 1/m)^m$$

and so $\mathbf{P}(T > 1/m) = \mathbf{P}(T > 1)^{1/m}$. Letting $t = 1/m$ in the pre-last display, we have

$$\mathbf{P}(T > n/m) = \alpha^{n/m}.$$

Now, for $t > 0$, let $q_1 > q_2 > \dots$ be rational numbers with $\inf\{q_1, q_2, \dots\} = t$. Then

$$\mathbf{P}(T > t) = \mathbf{P}(\cup_k \{T > q_k\}) = \sup_k \mathbf{P}(T > q_k) = \sup_k \alpha^{q_k} = \alpha^{\inf_k q_k} = \alpha^t.$$

Since $\alpha < 1$, we have $\lambda := -\log \alpha > 0$. □

We say that T is exponential with parameter (rate) λ.

It is easy to see that

Lemma 5.11. *If T is exponential with rate λ then T has density*

$$f(t) = \lambda e^{-\lambda t}, \quad t \geq 0.$$

Also,

$$\mathbf{E} e^{\theta T} = \frac{\lambda}{\lambda - \theta},$$

defined for all $\theta < \lambda$, and

$$\mathbf{E}T = 1/\lambda, \quad \operatorname{var} T = 1/\lambda^2.$$

Proof Note that

$$\int_0^t f(s) ds = 1 - e^{-\lambda t},$$

showing that f is a density for T. The rest are trivial. □

Lemma 5.12. *Let $T_1, T_2, \ldots, T_d$ be independent exponential random variables with parameters $\lambda_1, \lambda_2, \ldots, \lambda_d$, respectively. Then $\min(T_1, \ldots, T_d)$ is exponential with parameter $\lambda_1 + \cdots + \lambda_d$.*

Proof

$$\mathbf{P}(\min(T_1, \ldots, T_d) > t) = \mathbf{P}(T_1 > t) \cdots \mathbf{P}(T_d > t).$$

□

Whereas an exponential is the natural analogue of a geometric, in that they are both memoryless, the former also enjoys the important scaling property:

If T is exponential with rate λ then, for any $c > 0$, cT is exponential with rate λ/c,

and this is obvious.

EXERCISE 52. Let $T_1, T_2, \ldots, T_d$ be independent exponential random variables all with rate 1. Show that

$$\text{law of } \max(T_1, \ldots, T_d) = \text{ law of } \left(T_1 + \frac{T_2}{2} + \cdots + \frac{T_d}{d}\right).$$

Another relation between geometric and exponential is the following: Let p be very small. Let X be geometric with parameter p. Consider a scaling of X by p, i.e. the random variable pX which takes values $p, 2p, 3p, \ldots$. Then the law of pX converges to an exponential law with rate 1:

Lemma 5.13 (Rényi)**.** *For X geometric with parameter p,*

$$\lim_{p \to 0} \mathbf{P}(pX > t) = e^{-t}, \quad t > 0.$$

Proof $\mathbf{P}(pX > t) = \mathbf{P}(X > t/p)$, and, since X is an integer,

$$\mathbf{P}(X > t/p) = \mathbf{P}(X \geq \lceil t/p \rceil) = (1-p)^{\lceil t/p \rceil + 1}.$$

Since $\lim_{n \to \infty}(1 + n^{-1})^n = e$, we have that the last expression converges to e^{-t} as $p \to 0$. □

References

[1] B. Fristedt & L. Gray *A Modern Approach to Probability Theory*, Birkhäuser, 1997.

[2] E. Hewitt & K. Stromberg *Real and Abstract Analysis*, Springer, 1965.

[3] A. N. Kolmogorov & S. V. Fomin, *Introductory Real Analysis*, Dover, 1975.

[4] D. Mumford, The dawning of the age of stochasticity.
`http://www.dam.brown.edu/people/mumford/Papers/Dawning.pdf`, 1999.

[5] D. Williams, *Probability with Martingales*, Cambridge, 1991.