SMSTC (2007/08)

Probability

www.smstc.ac.uk

Contents

SMSTC (2007/08)

Probability

Lecture 3: Random variables

Member of staff, Heriot-Watt University[a]

www.smstc.ac.uk

Contents

[a]

3.1 Random variables and their distributions

3.1.1 Review of definitions

Recall that a RANDOM VARIABLE is a measurable function from a measurable space $(\Omega, \mathscr{F})$ into $(\mathbb{R}, \mathscr{B})$. More generally, A RANDOM ELEMENT IS A MEASURABLE FUNCTION X between two measurable spaces, an "abstract" one, $(\Omega, \mathscr{F})$, and a "concrete"[b] one, $(S, \mathscr{S})$. 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

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

For example, if $(S, \mathscr{S}) = (\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$, where $\mathscr{B}(\mathbb{R}^d)$ are the Borel sets on $\mathbb{R}^d$, we refer to X as d (REAL) RANDOM VARIABLES or a RANDOM VECTOR because we may, by choosing Cartesian

[a] These notes contain almost no proofs. For a complete set of notes, see [3]; alternatively, read the introductory book [4], or the more advanced book [1].

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

coordinates on $\mathbb{R}^d$, represent X by $(X_1, \dots, X_d)$, where X_1 is one random variable, X_2 is one random variable, ..., X_d is one random variable..

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

Recall that the σ-ALGEBRA GENERATED BY A COLLECTION, say $\mathscr{A}$, of subsets of Ω, is defined as the intersection of all σ-algebras containing $\mathscr{A}$; it is denoted by $\sigma(\mathscr{A})$.

Lemma 3.1. *If $X : \Omega \to S$ is a function, $\mathscr{F}$ a σ-algebra on Ω and $\mathscr{S}$ a σ-algebra 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 $B \in \mathscr{C}$.*

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-algebra 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 $x \in \mathbb{R}$.*

If X_n is a sequence of random variables then $X_1 + X_2, X_1 \cdot X_2, X_1 \wedge X_2$ are all random variables. Also, recall that $\inf_j X_j$, $\underline{\lim}_j X_j, \sup_j X_j$, $\overline{\lim}_j X_j$ are random variables in $\mathbb{R} \cup \{+\infty, -\infty\}$. If $X : \Omega \to S$ is a function and $\mathscr{S}$ a fixed σ-algebra on S, the σ-ALGEBRA GENERATED BY X is defined by

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

Let us study random elements in a product $S_1 \times \cdots \times S_d$ of sets. Suppose that on each S_i we have a σ-algebra $\mathscr{S}_i \subset 2^{S_i}$. We first construct a natural σ-algebra 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 $(S_i, \mathscr{S}_i)$, $i = 1, \dots, d$. 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 3.2. *Let $X_i : (\Omega, \mathscr{F}) \to (S_i, \mathscr{S}_i)$, $i = 1, 2, \dots, d$, be random elements. Let $S = S_1 \times \cdots \times S_d$, $\mathscr{S} = \mathscr{S}_1 \otimes \cdots \otimes \mathscr{S}_d$. Then $(X_1, \dots, X_d) : (\Omega, \mathscr{F}) \to (S, \mathscr{S})$ is a random element.*

EXERCISE 3. Show that, if $X = \mathbf{1}_A$, the indicator of a set A, then $\sigma(\mathbf{1}_A) = \{\varnothing, A, A^c, \Omega\}$. Consider next two subsets A_1, A_2 of Ω and prove that the σ-algebra generated by $c_1 \mathbf{1}_{A_1} + c_2 \mathbf{1}_{A_2}$ is

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

EXERCISE 4. Consider the following partition of Ω:

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

Show that 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-algebra 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-algebra which is contained in $\sigma(\mathscr{C})$. Show that if the values b_i are distinct then $\sigma(X) = \sigma(\mathscr{C})$.

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, OFTEN WRITTEN AS $X \in \sigma(Y)$ if

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

Lemma 3.3. *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 $Y = H \circ X$.*

EXERCISE 6. 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 3.3 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 3.4. *Let $X : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B})$. Then there exists a sequence $X_1, X_2, \dots$ 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.*

The proof of Lemma 3.4 can be found in [3] and is based on defining

$$\tau_n(x) := 2^{-n} \lceil 2^n x \rceil \wedge n, \quad X_n(\omega) := \tau_n(X(\omega)),$$

and the observation that X_n converges to X.

EXERCISE 7. Let X be one random variable in $\mathbb{R}$. Show that $\sigma(X^2)$ is a strict subset of $\sigma(X)$. But show that $\sigma(2^X) = \sigma(X)$.

3.1.2 Law (or distribution) of a random element (or variable)

The reader will of course have noticed that the concept of "random element" has nothing to do with randomness.[c] Let $\mathbf{P}$ be a probability on $(\Omega, \mathscr{F})$. The DISTRIBUTION OR LAW of the random element X is a probability $\mathbf{P}_X$ on $(S, \mathscr{S})$ which is generated, 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 8. Show that if $(\Omega, \mathscr{F}, \mathbf{P})$ is a probability space and $X : (\Omega, \mathscr{F}) \to (S, \mathscr{S})$ a random element 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.

[c] The name is justified because of the way the concept is used.

EXERCISE 9. With the notation of Exercise 1, show that the law of $H \circ X$ as a random element on the probability space $(\Omega, \mathscr{F}, \mathbf{P})$ is the same as the law of H as a random element on $(S, \mathscr{S}, \mathbf{P}_X)$.

So the role of a random element is to transform an abstract probability space into a concrete one. In practise, one is often given[d] 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 $\mathbf{P}_X = \mathbf{Q}$. In the absence of any further requirement we 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, $\mathbf{P}_X = \mathbf{Q}$.

3.1.3 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 PROBABILITY MASS FUNCTION. 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 $\mathbf{P}_X$.

EXERCISE 10. 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.

3.1.4 Uniform random variable and continuous random variables

A uniform random variable U in the interval $[0, 1]$ is such that its law satisfies $\mathbf{P}_U([a, b]) = b - a$ for all $0 \le a \le b \le 1$. To show that such a random variable exists is hard and the reader is referred to [3], for an approach based on using the probability space $\Omega = \{0, 1\}^{\mathbb{N}}$. A more standard approach is to define $U(x) = x$, $x \in [0, 1]$, where $[0, 1]$ is endowed with the Borel σ-algebra and the Lebesgue measure $\mathbf{P}$. Having shown that the latter exists, we simply observe that $\mathbf{P}_U$ is the law of a uniform random variable. Notice that if D is a countable set then

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

For example, $\mathbf{P}(U \text{ is rational number }) = 0$.

If $F : \mathbb{R} \to \mathbb{R}$ is a nondecreasing continuous function with $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to \infty} F(x) = 1$, and if $F^{-1}(u) := \inf\{x \in \mathbb{R} : F(x) > u\}$ then $X = F^{-1}(U)$ is a random variable that shares something in common with U, namely,

$$\mathbf{P}(X \in D) = 0,$$

if D is a countable set. Such a random variable is called continuous.

[d] 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.

3.2 Distribution functions

The distribution function of a random variable X is defined by:

$$F(x) := \mathbf{P}_X(-\infty, x], \quad x \in \mathbb{R}. \tag{3.1}$$

This function is useful because:

Lemma 3.5. *Knowledge of* $\mathbf{P}_X$ *on this class of sets only (semi-infinite intervals) implies knowledge of* $\mathbf{P}_X$ *on the whole of* $\mathscr{B}$.

For a proof, see [3]. Next, here are some properties of F:

Lemma 3.6. *(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)* $\lim_{n\to\infty} F(x + 1/n) = F(x)$.

Remark:
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 3.6 would be replaced by $\lim_{n\to\infty} F(x - 1/n) = F(x)$.

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 3.2. *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* $x \in \mathbb{R}$.

Lemma 3.7. *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)* $\mathbf{P}(X = a) = F(a) - F(a-)$.

EXERCISE 11. Carefully justify the formulae in the proof of Lemma 3.7.

3.2.1 Types of distribution functions

We now discuss the various kinds of distribution functions on $\mathbb{R}$.

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 3.1. Let X be a random variable such that $\mathbf{P}(X = n) = 2^{-n}$, $n \in \mathbb{N}$. Then its distribution function looks like in Figure 3.1.

Example 3.2. 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.)

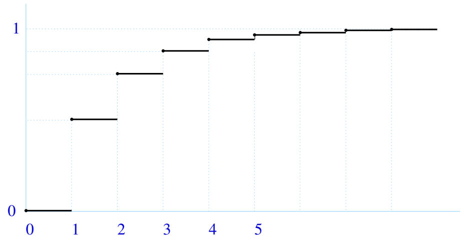

Figure 3.1: Distribution function of a discrete random variable

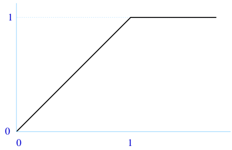

Figure 3.2: An absolutely continuous distribution function

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 3.3. 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 in Figure 3.2.

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.

EXERCISE 12. Consider the space $(\Omega = \{0, 1\}^{\mathbb{N}}, \mathscr{F}, \mathbf{P})$, where $\mathscr{F}$ is the product σ-algebra, and $\mathbf{P}$ is such that $\mathbf{P}\{\omega \in \Omega : \omega_1 = i_1, \ldots, \omega_n = i_n) = 2^{-n}$, $i_1, \ldots, i_n \in \{0, 1\}$, $n \in \mathbb{N}$. 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, see [4]. For the time being, you may think of it as the standard Riemann integral of Integral Calculus.

[f] unconsciously

Show that the random variable V defined in Example 12 has a continuous but not absolutely continuous distribution function.

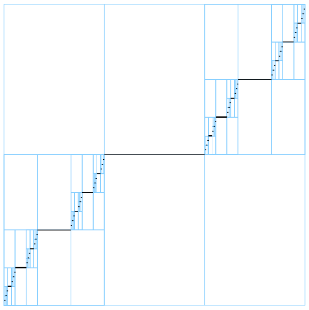

Figure 3.3: A continuous distribution function without density

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 3.1. *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].

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 $A^c \subset \cup_n I_n$. For further elaboration on this important notion, without which the concept of a density cannot be properly understood, see [2, 3].

EXERCISE 13. Show that a continuous and piecewise differentiable function G is almost everywhere differentiable.

3.3 Transformation rules and densities

Consider a random variable $X : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B})$. 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 $H : (\mathbb{R}, \mathscr{B}) \to (\mathbb{R}, \mathscr{B})$ 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 3.2. *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 14. Let U be a uniform random variable in the interval $(0, 1)$. Find the density function of e^{e^U}.

Theorem 3.2 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 15. 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 $\varphi(X)$.

3.4 Expectation

The expectation $\mathbf{E}X$ of one real random variable X is, if it can be defined, 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

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

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

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

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.

The general case requires an approximation procedure, known as integral with respect to a measure. When X is a nonnegative random variable, there always exists a sequence ξ_n of nonnegative random variables, each of which takes finitely many values, such that $\xi_n(\omega) \to X(\omega)$, as $n \to \infty$, for all $\omega \in \Omega$. We then define

$$\mathbf{E}X = \lim_{n\to\infty} \mathbf{E}\xi_n.$$

It can be proved [3] that the limit exists and is independent of the approximation procedure. If X has no restriction on sign, we define

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

provided that not both terms are $+\infty$.

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 3.8 (algebraic properties of expectation). *Suppose X, Y are integrable random variables on the same probability space $(\Omega, \mathscr{F}, \mathbf{P})$. Then:*
(i) If $\mathbf{P}(A) = 0$ then $\mathbf{E}X\mathbf{1}_A = 0$.
(ii) $\mathbf{E}(cX) = c\mathbf{E}(X)$ for all $c \in \mathbb{R}$.
(iii) $\mathbf{E}(X + Y) = \mathbf{E}X + \mathbf{E}Y$.
(iv) If $\mathbf{P}(X \geq 0) = 1$ then $\mathbf{E}X \geq 0$. If $\mathbf{P}(X \leq Y) = 1$ then $\mathbf{E}X \leq \mathbf{E}Y$.

Theorem 3.3 (monotone convergence theorem). *For ANY sequence X_n of nonnegative random variables such that $X_n \uparrow X$, we have $\mathbf{E}X_n \uparrow \mathbf{E}X$.*

Lemma 3.9 (Fatou's lemma)**.** *For any sequence X_n of nonnegative random variables,* $\mathbf{E}\,\underline{\lim}\,X_n \le \underline{\lim}\,\mathbf{E}X_n$.

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

Theorem 3.4 (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.*

3.4.1 Substitution rule

There is a little problem with the notation $\mathbf{E}X$ in that it uses the same letter regardless of the probability $\mathbf{P}$. More correctly, one should write $\mathbf{E}_P X$ instead of $\mathbf{E}X$ or $\int_\Omega X d\mathbf{P}$. Let us consider compositions of random elements.

Lemma 3.10. *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 either 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 $(S, \mathscr{S}, \mathbf{P}_Z)$.)

Sketch of 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 3.1.3). 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 general random variable. Use an approximation by simple random variables to conclude. □

Corollary 3.3 (the expectation of a random variable depends only on its law)**.** *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: $\iota(x) \equiv x$.

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 $\mathbf{E}X = \sum_{x \in S} x\mathbf{P}(X = x) = \sum_{x \in S} xp(x)$, 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 $\int_{\mathbb{R}} xf(x)dx$. To do this, requires understanding of Lebesgue-Stieltjes integral. The step is omitted and can be found in [1,4], and in [3].

Lemma 3.11. *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.

If you want to see the proof, look at [3]. But do notice that, 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 3.4. *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)

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

The proof of (i) follows from Lemma 3.11 with $g = \iota$. The proof of (ii) follows from corollary 3.3.

EXERCISE 17. 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 18. 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 19. 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.)

3.5 Inequalities

3.5.1 Markov, Chebyshev, Chernoff

Lemma 3.12 (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 3.2. 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 3.13 (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 3.14 (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 20. 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.

3.5.2 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 3.15. *Let X be a real integrable random variable and φ a convex function. Then*

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

EXERCISE 21. 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 $A_n \geq G_n \geq H_n$.

3.6 Moments

Definition 3.3. 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 3.16. *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 3.5. *If $\mathbf{E}|X|^p < \infty$ for some $p > 0$ then $\mathbf{E}|X|^r < \infty$ for all $0 < r < p$.*

3.7 Hölder, Minkowski and Cauchy-Bunyakowskii-Schwarz

Definition 3.4. 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 $\mathrm{cov}(X, Y) := \mathbf{E}((X - \mathbf{E}X)(Y - \mathbf{E}Y))$ is called COVARIANCE between X and Y.

Lemma 3.17 (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 $W : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B})$,

$$\mathbf{E}_q W = \frac{\mathbf{E}(Y^q W)}{\mathbf{E}(Y^q)}. \tag{3.2}$$

Letting, in (3.2), $W = XY^{1-q}$, 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 3.16 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 3.6. (Cauchy-Bunyakowskii-Schwarz) *Let X, Y be real random variables. Then*

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

as long as all terms involved exist and are finite.

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

Corollary 3.7. *Let X, Y be real random variables. Let*

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

whenever the terms exist. Then

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

Lemma 3.18 (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:

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

□

3.8 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 3.10) 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 3.19. *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 3.4, 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. □

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] T. Konstantopoulos, Extended set of lecture notes, with proofs,
`www.ma.hw.ac.uk/~takis`

[4] D. Williams, *Probability with Martingales*, Cambridge, 1991.