CHAPTER 2. ELEMENTS OF PROBABILITY

Probability Theory is the branch of Analysis that deals with the study of random phenomena. As we shall see in the next section, it has its own vocabulary and notation. This is due partly to the circumstances of the subject's development – starting in the 16^{th} and 17^{th} centuries, long before the connection with measure and integration was made explicit by Kolmogorov in the early 1930's – and partly to the fact that the central problems, motivation, ideas, techniques, and intuitive content of Probability are distinctly its own. In this chapter we shall discuss only the very basic aspects of this subject, including notions such as independence and conditional expectations and results such as the Law of Large Numbers, the Central Limit Theorem and Cramér's Theorem on "Large Deviations". We shall continue the development of this subject in subsequent chapters.

2.1. PROBABILITY SPACES AND RANDOM VARIABLES

A *probability space* is a measure space $(\Omega, \mathcal{F}, \mathbf{P})$ with total mass $\mathbf{P}(\Omega) = 1$. The set Ω is now called the "sample space"; it can be thought of as representing the collection of all possible outcomes of a random experiment, whose evolution is not possible to predict in advance with certainty. The subsets of Ω in the σ-algebra $\mathcal{F}$ are called "events", and the measure $\mathbf{P}$ assigns to each of them a number in the interval $[0, 1]$ that represents the "probability of its occurrence".

Within this framework, a *random vector* is just a measurable, real-valued function $X : \Omega \to \mathbf{R}^d$; if $d = 1$, we say that X is a **random variable.** The quantity $X(\omega)$ represents a vector of numerical characteristics assigned to the outcome $\omega \in \Omega$ of our random experiment, that we happen to be interested in. The probability measure "induced by X on the Borel subsets of $\mathbf{R}^d$", namely

$$\mu_X(B) := \mathbf{P}\big(\{\omega \in \Omega \,|\, X(\omega) \in B\}\big) \equiv \mathbf{P}(X \in B) = (\mathbf{P} \circ X^{-1})(B)\,, \quad B \in \mathcal{B}(\mathbf{R}^d)\,, \tag{1.1}$$

is often called the *distribution of the random vector* X. This measure generates the *probability distribution function*

$$F_X(x) := \mu_X((-\infty, x]) = \mathbf{P}[X \leq x]\,, \quad x \in \mathbf{R}^d \tag{1.2}$$

of the random vector X, in the notation of §1.4.C. It is clear that this function $F_X : \mathbf{R}^d \to [0, 1]$ satisfies the conditions of Definition 1.4.2 (or of Definition 1.4.1, in the case of $d = 1$). We say that two random vectors X and Y are *identically distributed*, if they induce the same measures ($\mu_X = \mu_Y$ in the notation of (1.1)) on the on the Borel subsets of $\mathbf{R}^d$.

1.1 Proposition : Skorohod Construction. *For any given probability distribution function* $F : \mathbf{R}^d \to [0,1]$*, there exists a probability space* $(\Omega, \mathcal{F}, \mathbf{P})$ *and a random vector* $X : \Omega \to \mathbf{R}^d$*, such that* $F_X(\cdot) \equiv F(\cdot)$*. In the special case* $d = 1$*, the mappings*

$$X^+(\omega) := \inf\{x \,|\, F(x) > \omega\}\,, \quad X^-(\omega) := \inf\{x \,|\, F(x) \geq \omega\}\,, \qquad 0 \leq \omega \leq 1 \qquad (1.3)$$

or equivalently

$$X^+(\omega) = \sup\{x \,|\, F(x) \leq \omega\}\,, \quad X^-(\omega) := \sup\{x \,|\, F(x) < \omega\}\,,$$

on the probability space $(\Omega, \mathcal{F}, \mathbf{P}) = ([0,1], \mathcal{B}([0,1]), \lambda)$ *are random variables with probability distribution functions* $F_{X^\pm}(\cdot) \equiv F(\cdot)$*.*

Proof: An obvious choice is to take X the identity mapping on the space $(\mathbf{R}^d, \mathcal{B}(\mathbf{R}^d), \mu_F)$ where μ_F is the Lebesgue-Stieltjes measure of Section 1.4 corresponding to F, and note $F_X(x) = \mathbf{P}(X \in (-\infty, x]) = \mu_F((-\infty, x]) = F(x),\ \forall\ x \in \mathbf{R}^d$.

For the one-dimensional case $d = 1$, there is another, very useful, construction due to A.V. Skorohod. One looks at the right- and left-continuous inverses of F, namely the mappings $X^\pm$ of (1.3) on the space $(\Omega, \mathcal{F}, \mathbf{P}) \equiv ([0,1], \mathcal{B}([0,1]), \lambda)$ (see Figure), and notices the implications $\omega \leq F(x) \Leftrightarrow X^-(\omega) \leq x$, $\omega < F(x) \Rightarrow X^+(\omega) \leq x$, valid for all $\omega \in [0,1]$, $x \in \mathbf{R}$. This implies

$$\mathbf{P}[X^- \leq x] = \lambda\big(\{\omega \,|\, \omega \leq F(x)\}\big) = F(x) \leq \mathbf{P}[X^+ \leq x]\,, \quad \forall\ x \in \mathbf{R}\,.$$

Of course $X^+ \geq X^-$, so $\{X^+ \neq X^-\} = \cup_{q \in \mathbf{Q}} \{X^- \leq q < X^+\}$. But we have

$$\mathbf{P}[X^- \leq q < X^+] \;=\; \mathbf{P}\left[\{X^- \leq q\} \setminus \{X^+ \leq q\}\right] \;=\; F(x) - \mathbf{P}[X^+ \leq x]\,, \quad \forall\ q \in \mathbf{R}\,,$$

and thus $\mathbf{P}[X^+ \neq X^-] = 0$, because $\mathbf{Q}$ is countable. Therefore, $\mathbf{P}[X^+ \leq x] = F(x)\,,\ \forall x \in \mathbf{R}\,.$ $\diamond$

For any random variable $Y \in \mathbf{L}^1$, the integral $\int_\Omega Y\, d\mathbf{P}$ is denoted by $\mathbf{E}(Y)$ and is called the **expectation** of Y. If $Y \in \mathbf{L}^k$ for some $k > 0$, the integrals $\mathbf{E}(Y^k) = \int_\Omega Y^k\, d\mathbf{P}$ and $\mathbf{E}(|Y|^k) = \int_\Omega |Y|^k\, d\mathbf{P}$ are called k^{th} **moment** and the k^{th} **absolute moment** of Y, respectively. If $Y \in \mathbf{L}^2$, the non-negative quantity

$$\mathrm{Var}(Y) := \mathbf{E}\Big(Y - \mathbf{E}(Y)\Big)^2 = \mathbf{E}(Y^2) - \big(\mathbf{E}(Y)\big)^2$$

is called the **variance** of Y; it vanishes if and only if Y is "degenerate", in the sense $\mathbf{P}[Y = y] = 1$, for some $y \in \mathbf{R}$. The square-root $\sqrt{\mathrm{Var}(Y)}$ of the variance is called *standard deviation.* If $Y\,, Z$ are two random variables in $\mathbf{L}^2$, we define their **covariance**

$$\mathrm{Cov}(Y, Z) := \mathbf{E}\Big[(Y - \mathbf{E}(Y)) \cdot (Z - \mathbf{E}(Z))\Big] = \mathbf{E}(YZ) - \mathbf{E}(Y)\,\mathbf{E}(Z)\,.$$

We say that the variables are **uncorrelated**, if their covariance is zero.

1.2 Proposition: *If X is a random vector with values in $\mathbf{R}^d$ and $h : \mathbf{R}^d \to \mathbf{R}$ a measurable function, then $h(X)$ is a random variable; and if this random variable is integrable, its expectation can be written in terms of the distribution of the random vector X and in the notation of* (1.4.2)$'$, *as*

$$\mathbf{E}[h(X)] = \int_{\mathbf{R}^d} h \, d\mu_X = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(x_1, \cdots, x_d) \, dF(x_1, \cdots, x_d) \, . \tag{1.4}$$

Proof : This is obvious from the definition of μ_X , if $h = \chi_E$ is the indicator function of some Borel set $E \in \mathcal{B}(\mathbf{R}^d)$. Thus (1.4) is also valid for simple functions; and by the usual approximation technique, also for nonnegative, as well for μ_X-integrable, functions.

1.1 Remark: For any given $\xi \in \mathbf{R}^d$ we may choose the real and imaginary parts of the function $h(x) := \exp(i\langle \xi, x \rangle) = \exp(i \sum_{j=1}^d \xi_j x_j)$ $x \in \mathbf{R}^d$ where $i = \sqrt{-1}$. Then the function $\varphi_X : \mathbf{R}^d \to \mathbf{C}$ defined by

$$\varphi_X(\xi) := \mathbf{E}\left[e^{\,i\langle \xi, X \rangle} \right] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{\,i\langle \xi, x \rangle} \, dF(x_1, \cdots, x_d)$$

is called the *characteristic function* of the random vector X. We shall see in Chapter 3 that $\varphi_X(\cdot)$ determines uniquely the distribution function $F(\cdot)$.

1.1 Definition: A probability distribution function $F(\cdot)$ (as well as a random variable Y with $F_Y(\cdot) \equiv F(\cdot)$) is called

(i) *absolutely continuous*, if it is of the form $F(x) = \int_{-\infty}^x f(u) \, du$, $x \in \mathbf{R}$ for some meassurable “probability density-function” $f : \mathbf{R} \to [0, \infty)$ with $\int_{-\infty}^{\infty} f(u) \, du = 1$;

(ii) *purely discrete*, if it is of the form $F(x) = \sum_{\substack{k \in \mathcal{K} \\ u \le x}} p(k)$, $x \in \mathbf{R}$, for some finite or countably infinite set $\mathcal{K} \subset \mathbf{R}$ and a “probability mass-function” $p : \mathcal{K} \to [0, \infty)$ with $\sum_{k \in \mathcal{K}} p(k) = 1$. (Compare with Exercise 1.7.2.)

1.1 Example: The following are examples of absolutely continuous distributions (with densities $f : \mathbf{R} \to [0, \infty)$ as listed):

- *Exponential:* $f(x) = \lambda \, e^{-\lambda x} \chi_{(0,\infty)}(x)$, for some $\lambda > 0$.
- *Gamma* $\Gamma(\lambda, r)$: $f(x) = \frac{\lambda^r}{\Gamma(r)} \, x^{r-1} \, e^{-\lambda x} \, \chi_{(0,\infty)}(x)$, with parameters $\lambda > 0$ and $r > 0$.

Here $\Gamma(r) = \int_0^{\infty} x^{r-1} \, e^{-x} \, dx$ is the Gamma function. The exponential is a special case of this distribution, corresponding to $r = 1$.

- *Standard Normal* $\mathcal{N}(0\,,1)$*:* $f(x) = \frac{1}{\sqrt{2\,\pi}} \, e^{-x^2/2}$.
- *Normal* $\mathcal{N}(m\,,\sigma^2)$*:* $f(x) = \frac{1}{\sqrt{2\,\pi\,\sigma^2}} \, e^{-(x-m)^2\,/\,2\sigma^2}$, with parameters $m \in \mathbf{R}$ and $\sigma^2 > 0$.

- *Uniform on* $[0,h]$*:* $f(x) = \frac{1}{h}\,\chi_{[0,h]}(x)$, for some $h > 0$.
- *Uniform on* $[-h,h]$*:* $f(x) = \frac{1}{2\,h}\,\chi_{[-h,h]}(x)$, for some $h > 0$.
- *Double Exponential:* $f(x) = (1/2)\,e^{-|x|}$.
- *Cauchy:* $f(x) = \big(\pi\,(1+x^2)\big)^{-1}$.
- *Triangular on* $[-1,1]$*:* $f(x) = (1-|x|)\,\chi_{[-1,1]}(x)$.
- *Fejér:* $f(x) = \frac{1-cos\,x}{\pi\,x^2}$.
- *Logistic:* $f(x) = e^{-x}/(1+e^{-x})^2$.

1.2 Example: The following are examples of purely discrete distributions (with mass-functions $p(\cdot)$ as listed):

- *Dirac measure* δ_a *:* $p(a) = 1$ for some $a \in \mathbf{R}$.
- *Bernoulli:* $p(a) = p$, $p(b) = 1 - p =: q$ for some $p \in (0,1)$ and $a\,, b \in \mathbf{R}$.
- *Symmetric Bernoulli:* $p(b) = p(-b) = 1/2$, for some $b \in (0,\infty)$.
- *Binomial* $\mathcal{B}(n,p)$ *:* $p(k) = (n!\,/k!\,(n-k)!) \cdot p^k\,(1-p)^{n-k}$, $k = 0, \cdots, n$,
  with parameters $n \in \mathbf{N}$ and $p \in (0,1)$.
- *Poisson:* $p(k) = e^{-\lambda}\,\frac{\lambda^k}{k!}$, $k = 0, 1, \cdots$, for some $\lambda > 0$.
- *Geometric:* $p(k) = p\,(1-p)^{k-1}$, $k = 1, 2, \cdots$, for some $p \in (0,1)$.

1.1 Exercise: Show by example, that (very!) different random variables can have the same distribution.

1.2 Exercise : Layered representation of the expectation. For any random variable $Z : \Omega \to [0,\infty)$, we have

$$\mathbf{E}(Z) = \int_0^\infty \mathbf{P}(Z > u)\,du \;\le\; \sum_{n=0}^{\infty} \mathbf{P}(Z > n) \;\le\; 1 + \mathbf{E}(Z)\,. \tag{1.5}$$

(*Hint:* Write $Z = \int_0^\infty \chi_{\{Z>u\}}\,du$, and then use Fubini-Tonelli; recall Exercise 1.6.4.)

1.3 Exercise: For the distributions of Examples 1.1 and 1.2, compute the expectation, variance, and moments of all orders, wherever these exist.

1.4 Exercise: Show that the distribution $\mu_X(\cdot) = \mathbf{P}(X \in \cdot)$ of a random vector $X : \Omega \to \mathbf{R}^d$ as in (1.1), is determined uniquely by knowledge of the expectations

$$\mathbf{E}[\Phi(X)] = \int_\Omega \Phi(X(\omega))\,d\mathbf{P}(\omega) = \int_{\mathbf{R}^d} \Phi\,d\mu_X$$

for *all* bounded, continuous functions $\Phi : \mathbf{R}^d \to \mathbf{R}$.

1.5 Exercise: On a given probability space, let $\mathcal{G}$ be a subspace of the Hilbert space $\mathbf{L}^2(\mathbf{P})$ whose closure does *not* contain the constant 1, and denote by $\pi : \mathbf{L}^2(\mathbf{P}) \to \mathcal{G}^\perp$ the projection-operator on $\mathcal{G}^\perp := \big\{ Z \in \mathbf{L}^2(\mathbf{P}) \,|\, \mathbf{E}(Zg) = 0\,,\ \forall\, g \in \mathcal{G} \big\}$, the orthogonal complement of $\mathcal{G}$, namely:

$$\mathbf{E}\left[(H - \pi(H)) \cdot Z \right] = 0\,, \qquad \text{for all} \quad H \in \mathbf{L}^2(\mathbf{P})\,,\ Z \in \mathcal{G}^\perp$$

(recall Theorem B.1, Appendix B).

(i) Show that $0 < \mathbf{E}[\pi^2(1)] = \mathbf{E}[\pi(1)] \le 1$.

(ii) Establish the inequality

$$\mathbf{E}[\pi^2(H)] \ge \frac{(\,\mathbf{E}[\pi(H)]\,)^2}{\mathbf{E}[\pi(1)]}\,, \qquad \forall\, H \in \mathbf{L}^2(\mathbf{P})\,.$$

(iii) Show that $\tilde{D} := \pi(1) / \mathbf{E}[\pi(1)]$ is the *minimum-variance* element of the (closed and convex) set $\mathcal{D} := \big\{ D \in \mathcal{G}^\perp \,|\, \mathbf{E}(D) = 1 \big\}$.

1.6 Exercise: Inclusion-Exclusion Formulae. Let $E_1, E_2, \cdots$ be arbitrary events on a probability space. Show that we have

$$\mathbf{P}\left(\cup_{i=1}^n E_i \right) = \sum_{i=1}^n \mathbf{P}(E_i) - \sum_{i<j} \mathbf{P}(E_i \cap E_j) + \sum_{i<j<k} \mathbf{P}(E_i \cap E_j \cap E_k) - \cdots + (-1)^n\, \mathbf{P}\left(\cap_{i=1}^n E_i \right).$$

(*Hint:* Argue that $\chi_{\cup_{i=1}^n E_i} = 1 - \prod_{i=1}^n (1 - \chi_{E_i})$; then expand the right-hand side and take expectations.)

1.7 Exercise: Bonferroni Inequalities. Let $E_1, E_2, \cdots$ be arbitrary events on a probability space. Show that we have

$$\mathbf{P}\left(\cup_{i=1}^n E_i \right) \le \sum_{i=1}^n \mathbf{P}(E_i)$$

$$\mathbf{P}\left(\cup_{i=1}^n E_i \right) \ge \sum_{i=1}^n \mathbf{P}(E_i) - \sum_{i<j} \mathbf{P}(E_i \cap E_j)$$

$$\mathbf{P}\left(\cup_{i=1}^n E_i \right) \le \sum_{i=1}^n \mathbf{P}(E_i) - \sum_{i<j} \mathbf{P}(E_i \cap E_j) + \sum_{i<j<k} \mathbf{P}(E_i \cap E_j \cap E_k)$$

and so on: if we stop the inclusion/exclusion formula of the right-hand side after an even (odd) number of steps, we get a lower (upper) bound.

1.8 Exercise: Suppose the real-valued random variables X, Y are identically distributed and we have $X \ge Y$ a.e. Show then $X = Y$ a.e.