1. Random Variables

A **probability space** (abbreviated "p.s.") (Ω, F, P) is a measure space such that

$$P(\Omega) = 1,$$

i.e. the total measure is 1. The set Ω ($\neq \emptyset$) is the **sample space** (abbreviated "s.s."), the members of the σ-algebra F are the **events**, and P is the **probability** (**measure**). Although Ω is always in the background, we rarely refer directly to it. Usually Ω is a huge set, and hence the class of events F cannot contain all subsets of Ω. However, since F is a σ-algebra, it follows that countable operations of events are events.

If a property holds for almost every ω (with respect to the measure P), namely for all $\omega \in \Omega \backslash N$, where $P(N) = 0$, we say that it holds **almost surely** (abbreviated "a.s.", or "P-a.s.", in the presence of more than one probability measures).

A real-valued function X on Ω,

$$X : \Omega \to \mathbb{R},$$

which is measurable with respect to F (symbolically $X \in F$, meaning that $\{X \leq x\} \stackrel{\text{def}}{=} \{\omega \in \Omega : X(\omega) \leq x\}$ is in F, for all $x \in \mathbb{R}$) is called **random variable**, abbreviated as "r.v.". Thus, if X is an r.v. and $x \in \mathbb{R}$, then $\{X \leq x\}$ is an event. It follows that $\{X \in B\}$ is an event, for every Borel set $B \subset \mathbb{R}$. Sometimes it is necessary to allow X to be an **extended real-valued function**, namely $X : \Omega \to \overline{\mathbb{R}} = [-\infty, \infty]$, but this does not really cause any serious complications.

The (**probability**) **distribution function** (abbreviated "d.f.") F of an r.v. X is defined by

$$F(x) = P(X \leq x), \qquad x \in \mathbb{R}. \tag{1}$$

Notice that F is increasing, $F(-\infty) = 0$, $F(\infty) = 1$, and F is right continuous, i.e. $F(x+) = F(x)$, since P is a measure. Thus, X induces a (probability) measure (abbreviated "p.m.") ν_X on the Borel sets of $\mathbb{R}$ so that

$$\nu_X\{(-\infty, x]\} = F(x) = P(X \leq x), \qquad x \in \mathbb{R}.$$

Hence, for every Borel set $B \subset \mathbb{R}$ we have

$$\nu_X(B) = P(X \in B).$$

The defining equation (1) of F can be also written in a Stieltjes integral form

$$P(X \leq x) = \int_{-\infty}^{x} dF(\xi), \qquad F(-\infty) = 0.$$

If F is absolutely continuous with $F'(x) = f(x)$, then

$$P(X \leq x) = \int_{-\infty}^{x} f(\xi) d\xi,$$

and f is the (**probability**) **density function** (abbreviated "p.d.f.") of X. All the statistical information (the law) of X, is contained in F (and hence, in f, if it exists).

Exercise 1. Consider the probability space (W, B, μ), where the sample space W is the open interval $(0,1)$, B is the σ-algebra of the Borel subsets of $(0,1)$, and μ is the Lebesgue measure (so that $d\mu = dx$). Given an arbitrary d.f. $F(x)$, find a r.v. Y on (W, B, μ) which is also an increasing function on $(0,1)$, such that the d.f. of Y is $F(x)$. Is Y unique? $(Ans.\ Y(w) = \sup\{y \in \mathbb{R} : F(y) \leq w\})$.

Exercise 2. For any d.f. $F(x)$ and any $a \in \mathbb{R}^+$ we have

$$\int_{-\infty}^{\infty} [F(x+a) - F(x)]\, dx = a.$$

Let $X : \Omega \to \mathbb{R}$ be any function and B the σ-algebra of the Borel subsets of $\mathbb{R}$. Then the set

$$\sigma(X) = \left\{X^{-1}(B) : B \in B\right\} \tag{2}$$

is a σ-algebra of subsets of Ω. In fact, it is the smallest σ-algebra with respect to which X is measurable; it is called the **σ-algebra generated by** X. Obviously X is an r.v. if and only if $\sigma(X) \subset F$.

If we consider a bunch of r.v.'s, say $X_1, ..., X_m$ (equivalently a **random vector** $X = (X_1, ..., X_m) : \Omega \to \mathbb{R}^m$) then the individual distributions of every X_j alone are not enough to determine their statistical properties as a group (i.e. as a random vector), since we also need to know how they are related (e.g., $X_1 =$ the I.Q. of a person, $X_2 =$ the salary of the person, $X_3 =$ his/her age, etc.). For this reason, we consider the **joint distribution function** of $X_1, ..., X_m$, namely

$$F(x_1, ..., x_m) = P\left(X_1 \leq x_1, ..., X_m \leq x_m\right), \qquad x_1, ..., x_m \in \mathbb{R}.$$

Alternatively, the r.v.'s $X_1, ..., X_m$ induce a measure ν on the Borel sets of $\mathbb{R}^m$:

$$\nu(B) = P\left\{(X_1, ..., X_m) \in B\right\}$$

(it is easy to see that, for any Borel subset B of $\mathbb{R}^m$, $\{(X_1, ..., X_m) \in B\}$ is an event). If there is a (Lebesgue measurable) function f such that

$$P\left\{(X_1, X_2, ..., X_m) \in B\right\} = \int_B f(x_1, x_2, ..., x_m) dx_1 dx_2 \cdots dx_m,$$

then f is called the **joint** (**probability**) **density function** (abbreviated as "p.d.f.") of $X_1, X_2, ..., X_m$. All the statistical information of $X_1, X_2, ..., X_m$, is contained in F (and hence in f, if it exists).

If

$$F(x_1, ..., x_m) = F_1(x_1) \cdots F_m(x_m), \qquad \text{for all } x_1, ..., x_m \in \mathbb{R},$$

where F_j is the d.f. of X_j, $j = 1, ..., m$, then the r.v.'s $X_1, ..., X_m$ are (**totally**) **independent**. In this case, the induced measure ν is a product measure on $\mathbb{R}^m$. The notion of independence extends to any collection of r.v.'s. The r.v.'s X_i, $i \in I$, where I is any set of indices, are independent if, for any finite subset J of I, the r.v.'s X_i, $i \in J$, are independent.

2. Expected Value

We start with a rather sketchy discussion on how to formulate a general definition of the mean value of an r.v. $X : \Omega \to \mathbb{R}$. Of course, the mean value of X, if it can be defined, it has to be a real number (or $\pm\infty$). Some of the symbols often used to denote it are:

$$E[X], \quad \int_\Omega X dP, \quad \int_\Omega X(\omega) dP(\omega), \quad \int_\Omega X(\omega) P(d\omega)$$

(sometimes Ω is omitted under the integral sign). In the literature one can encounter various other names for the mean value of an r.v. X, such as **expectation** or **expected value** or **mean** or, sometimes, **average** (**value**) of X. In measure theoretic terms, it is the integral of X (over Ω) with respect to the measure P.

Step 1. Let A be an event (i.e. $A \in F$) and

$$X = \mathbf{1}_A,$$

the **indicator** (or, sometimes, **characteristic**) **function** of A, meaning that

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

It follows that X is a Bernoulli r.v. and

$$X = \begin{cases} 1, & \text{with probability } P(A)\,; \\ 0, & \text{with probability } 1 - P(A)\,. \end{cases}$$

Hence, its mean value must be $P(A)$. Therefore

$$E[\mathbf{1}_A] = \int_\Omega \mathbf{1}_A dP = \int_A dP = P(A). \tag{3}$$

Step 2. An r.v. is called **simple** if it is a (finite) linear combination of indicator functions. Thus, if X is simple, then it can be written as

$$X = \sum_{j=1}^m c_j \mathbf{1}_{E_j}, \tag{4}$$

where $E_j \in F$ and $c_j \in \mathbb{R}$, for any $j = 1, ..., m$. Since the expectation (i.e. the mean value) has to be a linear operation, we can use (3) of Step 1 and set

$$E[X] = \int_\Omega X dP = \sum_{j=1}^m c_j \int_\Omega \mathbf{1}_{E_j} dP = \sum_{j=1}^m c_j P(E_j)\,. \tag{5}$$

However, in order for $E[X]$ to be meaningful (well-defined), we need to make sure that it is independent of the representation of X. That is, if

$$X = \sum_{j=1}^k b_j \mathbf{1}_{B_j},$$

is any other way to express X of (4), then we must have

$$\sum_{j=1}^k b_j P(B_j) = \sum_{j=1}^m c_j P(E_j)\,.$$

We can establish the above equality by induction on m. First one should show that X of (5) has a (unique) **canonical representation**

$$X = \sum_{j=1}^{n} a_j \mathbf{1}_{A_j},$$

where $\{a_1, ..., a_n\}$ is the range of X (the fact that the range is a finite set requires proof) and $A_j = \{X = a_j\} = \{\omega \in \Omega : X(\omega) = a_j\}$, $j = 1, ..., n$ (notice that $\{A_1, ..., A_n\}$ is a partition of Ω). Then it is not hard to show that

$$\sum_{j=1}^{m} c_j P(E_j) = \sum_{j=1}^{n} a_j P(A_j),$$

which in turn implies that $E[X]$ of (5) is well-defined for any simple r.v. X (the novice is urged to first "play" with the case $m = 2$, i.e. with $X = c_1 \mathbf{1}_{E_1} + c_2 \mathbf{1}_{E_2}$).

Step 3. Now if X is an r.v. such that $X(\omega) \geq 0$, for all $\omega \in \Omega$, we define

$$E[X] = \int_\Omega X dP = \sup Y \text{ simple}, 0 \leq Y \leq X \int_\Omega Y dP. \tag{6}$$

This definition makes sense even if $X : \Omega \to [0, \infty]$. If $P(X = \infty) > 0$, then, of course, $E[X] = \infty$. But it can also happen that $E[X] = \infty$, even if $X < \infty$ a.s.

Given the r.v. $X \geq 0$, with $X < \infty$ a.s., in order to understand the above definition (6) a little more, let us introduce the simple r.v.'s

$$X_n = \sum_{k=0}^{n2^n} \left(\frac{k}{2^n} \right) \mathbf{1}_{E_{n,k}}, \tag{7}$$

where, for each $n = 1, 2, 3, ...$, the $E_{n,k}$'s are the events

$$E_{n,k} = \left\{ \omega \in \Omega : \frac{k}{2^n} \leq X(\omega) < \frac{k+1}{2^n} \right\}, \qquad 0 \leq k \leq n2^n,$$

i.e. the $E_{n,k}$'s define a partition P_n of the range of X, up to the level $E_{n,k}$'s . The reason for using binary fractions (i.e $k/2^n$) as partition points is to guarantee that P_{n+1} is a refinement (and extension) of P_n, and hence

$$X_n \leq X_{n+1}.$$

Next observe that (5) implies

$$E[X_n] = \int_\Omega X_n dP = \sum_{k=0}^{n2^n} \left(\frac{k}{2^n} \right) P(E_{n,k}) = \sum_{k=0}^{n2^n} \left(\frac{k}{2^n} \right) P\left\{ \frac{k}{2^n} \leq X < \frac{k+1}{2^n} \right\}. \tag{8}$$

Also,

$$E[X_n] \leq E[X_{n+1}],$$

which implies that

$$\sup_n E[X_n] = \lim_n E[X_n].$$

The following theorem gives a more precise way to, actually, define $E[X]$:

Theorem 1. Let $X \geq 0$ be an r.v. which is finite a.s., and X_n as in (7). Then

$$E[X] = \int_\Omega X dP = \lim_n E[X_n] = \lim_n \sum_{k=0}^{n2^n} \left(\frac{k}{2^n}\right) P\left\{\frac{k}{2^n} \leq X < \frac{k+1}{2^n}\right\},$$

where $E[X]$ is the expectation of X, as defined by (6).

Notice that, in order to be in agreement with the definition of a simple r.v., we are not allowed to take the upper limit of the dummy variable k in the sum, in (7), to be ∞, especially when X is unbounded. Our choice to take $0 \leq k \leq n2^n$ does the job, since $n2^n/2^n = n \to \infty$.

Step 4. Finally, if $X : \Omega \to \mathbb{R}$ (or, more generally, $X : \Omega \to \overline{\mathbb{R}} \stackrel{\text{def}}{=} [-\infty, \infty]$) is an arbitrary r.v., we introduce the r.v.'s

$$X^+(\omega) = \max\{X(\omega), 0\} \qquad \text{and} \qquad X^-(\omega) = \max\{-X(\omega), 0\},$$

the positive and negative parts of X respectively. Observe that $X^+(\omega) \geq 0$, $X^-(\omega) \geq 0$ (thus $E[X^+]$ and $E[X^-]$ are defined) and X can be decomposed as

$$X(\omega) = X^+(\omega) - X^-(\omega).$$

It is therefore natural to define

$$E[X] = E[X^+] - E[X^-], \tag{9}$$

whenever it makes sense. Thus, (i) if $E[X^+] < \infty$ and $E[X^-] < \infty$, then $E[X]$ is a real number; (ii) if $E[X^+] = \infty$ and $E[X^-] < \infty$, then $E[X] = \infty$; (iii) if $E[X^+] < \infty$ and $E[X^-] = \infty$, then $E[X] = -\infty$; and, finally, (iv) if $E[X^+] = E[X^-] = \infty$, then $E[X]$ can not be defined. Since

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

we have that $E[X^+] < \infty$ and $E[X^-] < \infty$ imply $E[|X|] < \infty$ and conversely.

Remark. The above steps describe a standard measure theoretic procedure of defining the integral of a measurable function. Note that, the idea here is to partition the range of the integrand function rather than its domain, as is done in order to define the Riemann integral. It turns out that this measure theoretic definition behaves nicely, especially in the cases where one needs to interchange integral and limit, as we will see in the theorems to follow.

If $E[|X|^p] < \infty$ for some $p > 0$, we say that $X \in L_p(\Omega)$ or, sometimes, in order to avoid confusion, $X \in L_p(\Omega, F, P)$ (notice that $L_p(\Omega, F, P) \supset L_q(\Omega, F, P)$, whenever $p < q$). If $p \geq 1$, then $L_p(\Omega, F, P)$ is a Banach space with norm

$$\|X\|_p = (E[|X|^p])^{1/p}$$

and $L_p(\Omega, F, P)$ becomes a Hilbert space if and only if $p = 2$. Observe that $E[X]$ is a real number if and only if $X \in L_1(\Omega, F, P)$, in which case we say that X is **integrable**.

The **variance** of X is

$$V[X] \stackrel{\text{def}}{=} E\left[(X-\mu)^2\right] = E[X^2] - E[X]^2, \qquad \text{where } \mu = E[X] \tag{10}$$

(thus $V[X] \geq 0$ and $V[X] = 0$ if and only if $X(\omega) = \mu$ a.s.). Notice that, if $X \in L_2(\Omega, F, P)$, then $V[X] < \infty$.

The **covariance** of two r.v.'s X and Y is

$$Cov(X,Y) \stackrel{\text{def}}{=} E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - E[X]E[Y],$$

where $\mu_X = E[X]$ and $\mu_Y = E[Y]$. We remind the reader of the (Cauchy-Schwartz-type) inequality

$$Cov(X,Y)^2 \leq V[X] V[Y].$$

If A is an event and X an r.v., we define the integral of X over A as

$$E[X; A] \stackrel{\text{def}}{=} \int_A X dP = E[X \mathbf{1}_A]. \tag{11}$$

In the case where $P(A) = 0$, we have $\int_A X dP = 0$, even if $X = \infty$ on A (that is why sometimes people say that in measure theory $0 \cdot \infty = 0$).

***Exercise 3.** Let $X : \Omega \to \mathbb{R}$ be an r.v.. We set $A_m = \{|X| \leq m\}$, where m is a positive real number. If

$$E[|X|] = \infty,$$

and $p > 1$, does it follow that

$$\lim_{m \to \infty} \{E[|X|^p; A_m] - E[X; A_m]^p\} = \infty \tag{12}$$

(and, hence that $X \in L_p(\Omega, F, P)$ if and only if the limit in (12) is finite)?

Remark. The definition (10) of $V[X]$ assumes that $\mu = E[X]$ is a real number or, equivalently, that $E[|X|] < \infty$. Alternatively, we could define

$$\widetilde{V}[X] \stackrel{\text{def}}{=} \lim_{m \to \infty} \left\{ E\left[X^2; A_m\right] - E[X; A_m]^2 \right\}.$$

Of course, if $E[|X|] < \infty$, then $\widetilde{V}[X] = V[X]$. But, if $E[|X|] = \infty$ and the above exercise has an affirmative answer (at least for $p = 2$), then $\widetilde{V}[X] = \infty$ in spite of the fact that $V[X]$ of (10) is **not** meaningful.

2.1. **Basic Properties of the Expectation.** We give a list of properties of the expectation. Sometimes when the statement seems hard, the following approach helps to write a straightforward proof: First one establishes the statement for the case where the r.v.'s involved are indicator funcions of events. Then we prove the statement for simple r.v.'s, and finally we pass to the limit invoking (6) or (8).

P1. If $X \leq Y$ a.s., then

$$E[X] \leq E[Y],$$

provided $E[X]$ and $E[Y]$ exist.

P2 (Linearity)**.** If X, Y are r.v.'s and $a, b \in \mathbb{R}$, then

$$E[aX + bY] = aE[X] + bE[Y],$$

provided the right side is meaningful, namely not $\infty - \infty$ or $-\infty + \infty$.

P3 (Additivity over sets)**.** If $A, B \in F$, with $P(A \cap B) = 0$, then

$$\int_{A \cup B} X dP = \int_A X dP + \int_B X dP,$$

provided the right side is meaningful (just notice that $\mathbf{1}_{A\cup B} = \mathbf{1}_A + \mathbf{1}_B$ a.s.).

P4. If $a, b \in \mathbb{R}$, then

$$V[aX + b] = a^2 V[X].$$

P5. If $X, Y \in L_2(\Omega, F, P)$ are independent, then

$$E[XY] = E[X]E[Y], \qquad \text{equivalently } Cov(X, Y) = 0.$$

P6. If $X, Y \in L_2(\Omega, F, P)$, then

$$V[X + Y] = V[X] + V[Y] + 2Cov(X, Y)$$

(thus, if X, Y are independent, then $V[X + Y] = V[X] + V[Y]$).

P7 (Jensen's inequality)**.** If $\varphi : \mathbb{R} \to \mathbb{R}$ is a convex function, and X and $\varphi(X)$ are integrable r.v.'s, then

$$\varphi(E[X]) \le E[\varphi(X)].$$

We continue with three important theorems.

Theorem 2 (Fatou's Lemma)**.** If $X_n \ge 0$ a.s., for all $n = 1, 2, ...$, and $X_n \to X$ a.s., then

$$E[X] \le \liminf E[X_n].$$

For the proof see [2].

Remark. An alternative form of the theorem is the following: If $X_n \ge 0$ a.s., for all $n = 1, 2, ...$, then

$$E[\liminf X_n] \le \liminf E[X_n].$$

Theorem 3 (the Monotone Convergence Theorem, abbreviated MCT)**.** If $X_n \ge 0$ a.s., for all $n = 1, 2, ...$, $X_n \to X$ a.s., and $X_n \le X$ a.s., then

$$E[X] = \lim_n E[X_n].$$

Proof. Theorem 2 implies

$$E[X] \le \liminf E[X_n]. \tag{13}$$

However, $X_n \le X$ a.s., implies (see Property P1) that $E[X_n] \le E[X]$, for all n, and hence

$$\limsup E[X_n] \le E[X]. \tag{14}$$

The statement follows immediately by combining (13) and (14).

Remark. Most texts state the MCT using the stronger hypothesis that $X_n \le X_{n+1}$ a.s.

Theorem 4 (the Dominated Convergence Theorem, abbreviated DCT)**.** If $X_n \to X$ a.s., and $|X_n| \le Y$ a.s., with $E[Y] < \infty$, then

$$E[X] = \lim_n E[X_n].$$

Proof. Apply Theorem 2 to $Y + X_n$ and to $Y - X_n$.

If $Y = M$ a.s. (i.e. a positive constant), then we get the following:

Corollary (the Bounded Convergence Theorem, abbreviated BCT)**.** If $X_n \to X$ a.s., and $|X_n| \leq M$ a.s., then

$$E[X] = \lim_n E[X_n].$$

The next theorem states that the expectation, as defined above, agrees with the expectation as defined in the elementary probability courses.

Exercise 4. Show that the above theorems (Theorems 2, 3, 4) remain true if, instead of a sequence $\{X_n\}_{n\in\mathbb{N}}$, we have a family $\{X_t\}_{t\in[0,b)}$ of r.v.'s, such that $\lim_{t\nearrow b} X_t = X$ a.s. (here $0 < b \leq \infty$).

Theorem 5. Let $g : \mathbb{R} \to \mathbb{R}$ be a continuous function, and X an r.v., with d.f. $F(x)$, such that $g(X)$ is integrable (i.e. $E[g(X)]$ is a real number). Then

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)dF(x), \tag{15}$$

where the integration in the right side is done in the Riemann-Stieltjes sense. In particular, if $F(x)$ has a density $f(x) = F'(x)$, then (15) becomes the familiar

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$$

(notice that, if $g(x) = x$, we get the formula of elementary probability).

Proof (sketch). For

$$g(x) = \mathbf{1}_{(a,b]}(x)$$

one can easily check that (15) is true. Then it follows immediately that (15) is also true if g is a **step function**, namely it has the form

$$g(x) = \sum_{j=1}^{n} c_j \mathbf{1}_{(a_j,b_j]}(x),$$

where $a_j, b_j, c_j \in \mathbb{R}$, for $j = 1, ..., n$, while the intervals $(a_1, b_1], ..., (a_n, b_n]$ are disjoint. If $g(x) \geq 0$ is continuous, we approximate it from below by step functions, and establish (15) by invoking the MCT (Theorem 3). Finally, the general case follows by writing

$$g(x) = g^+(x) - g^-(x),$$

as usual.

The theorem can be generalized in two directions: (a) one can consider a $g : \mathbb{R} \to \mathbb{R}$ which is a Borel function; and (b) one can consider the r.v.'s $X_1, ..., X_m$ and a $g : \mathbb{R}^m \to \mathbb{R}$. E.g., if $X_1, ..., X_m$ have a joint density f, then

$$E[g(X_1, ..., X_m)] = \int \cdots \int_{\mathbb{R}^m} g(x_1, ..., x_m)f(x_1, ..., x_m)dx_1 \cdots dx_m.$$

Example 1. A **Gaussian** or **normal** r.v. X is by definition an r.v. with density

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2},$$

where μ and σ are real constants (if $\sigma = 0$, then $X(\omega) = \mu$ a.s.). It turns out that

$$E[X] = \mu \qquad \text{and} \qquad V[X] = \sigma^2 \ .$$

Thus, a normal distribution is completely determined by its expectation and variance. The normal distribution plays a fundamental role in probability theory.

EXAMPLE 2. The r.v.'s $X_1, X_2, ..., X_m$ are **jointly Gaussian** or **normal** if their joint density is

$$f(x_1, x_2, ..., x_m) = \frac{1}{(2\pi)^{m/2}\sqrt{D}} e^{-\frac{1}{2}Q[x-\mu]},$$

where $\mu = (\mu_1, \mu_2, ..., \mu_m)$ is a constant vector, $x = (x_1, x_2, ..., x_m)$, $Q[y]$ is the positive definite quadratic form

$$Q[y] = \sum_{i=1}^{m}\sum_{j=1}^{m} b_{ij} y_i y_j, \qquad b_{ij} = b_{ji},$$

and $D = \det(b_{ij})$.

It turns out that

$$E[X_j] = \mu_j \qquad \text{and} \qquad Cov(X_i, X_j) = \frac{D_{ij}}{D},$$

where D_{ij} is the minor of D corresponding to the entry b_{ij}. Thus the joint normal density is completely determined by $E[X_j] = \mu_j$, $1 \le j \le m$, and the **covariance matrix** $\{Cov(X_i, X_j)\}_{1\le i,j\le m}$ which, in fact, is the inverse of the matrix (b_{ij}). If $X_1, X_2, ..., X_m$ are jointly normal, then any linear combination

$$Y = c_1 X_1 + c_2 X_2 + \cdots + c_m X_m$$

is a normal random variable (in many cases we can assume without loss of generality that $E[X_j] = 0$, for all j).

Exercise 5. Let X be an r.v. with d.f. $F(x)$. If $X \ge 0$ a.s., then

$$E[X] = \int_0^\infty x\, dF(x) = \int_0^\infty [1 - F(x)]\, dx = \int_0^\infty P(X > x)\, dx = \int_0^\infty P(X \ge x)\, dx.$$

3. Conditional Expectation

As we will see, in many cases one needs to consider various σ-algebras of events (i.e. subalgebras of F). If X is an r.v. and A a σ-algebra (of events), then X is measurable w.r.t. A, symbolically $X \in A$, if

$$X^{-1}(B) \in A \qquad \text{for any Borel set } B \subset \mathbb{R}.$$

Equivalently, $X \in A$, if there is a dense set $S \subset \mathbb{R}$ such that $\{X \le x\} \in A$, for every $x \in S$.

If $X : \Omega \to \mathbb{R}$ is an arbitrary function, then the smaller σ-algebra with respect to which X is measurable is denoted by $\sigma(X)$. In other words $\sigma(X) = \{A \subset \Omega : A = X^{-1}(B)$, for some Borel set $B \subset \mathbb{R}\}$, i.e. $\sigma(X)$ is the σ-algebra generated by the sets $\{X \le x\}$, $x \in \mathbb{R}$ (or just $x \in S$, for some dense set $S \subset \mathbb{R}$). Of course, X is an r.v. if and only if $\sigma(X) \subset F$.

Theorem 6 (the Doob-Dynkin Lemma)**.** If $X : \Omega \to \mathbb{R}$ and $Y : \Omega \to \mathbb{R}$ are arbitrary functions, then

$$Y \in \sigma(X) \qquad \text{if and only if} \qquad Y = g(X),$$

where $g : \mathbb{R} \to \mathbb{R}$ is a Borel function (the statement can be extended to higher dimensions).

For the proof see [1], Sec. 9.1.

Intuitively, the theorem says that $Y \in \sigma(X)$ if X can catch all subtleties of Y.

Suppose now we have a subalgebra A of F and an r.v. X. Then X may not be measurable with respect to A and it is, therefore, natural to look for a $Y \in A$ which somehow best approximates X. This leads to the following definition:

Definition. Let X be in $L_1(\Omega, F, P)$ and A a subalgebra A of F. Then the r.v. Y is called the **conditional expectation of X relative to A**, symbolically $Y = E[X \mid A]$, if

(i) $Y \in A$;

(ii) for every $A \in A$ we have

$$\int_A Y dP = \int_A X dP.$$

If $A = \sigma(Z)$, for some r.v. Z, then we write $E[X \mid Z]$ instead of $E[X \mid \sigma(Z)]$ and $E[X \mid Z]$ is called the **conditional expectation of X given Z**. If $X = \mathbf{1}_B$, for some event $B \in F$, then, instead of $E[\mathbf{1}_B \mid A]$, we sometimes write $P[B \mid A]$, the **conditional probability of B relative to A**.

Of course, Theorem 6 implies that there is a Borel function $g : \mathbb{R} \to \mathbb{R}$ such that $E[X \mid Z] = g(Z)$.

Theorem 7. Given $X \in L_1(\Omega, F, P)$ and a σ-algebra A (of events), the conditional expectation $Y = E[X \mid A]$ exists and is unique a.s., namely if Y_1 and Y_2 are conditional expectation of X relative to A, then $Y_1 = Y_2$ a.s.

Proof. For any $A \in A$ we define

$$\lambda(A) = \int_A X dP.$$

It is not hard to see that $\lambda(\cdot)$ is a signed measure on A and, furthermore, that $\lambda << P$ (λ is absolutely continuous with respect to P) meaning that $\lambda(A) = 0$ whenever $P(A) = 0$. Hence, the Radon-Nikodym Theorem tells us that there exists a function $Y \in L_1(\Omega, A, P)$ such that

$$\lambda(A) = \int_A Y dP$$

and that this Y (which is called the Radon-Nikodym derivative of λ with respect to P and denoted by $d\lambda/dP$) is unique a.s. It follows immediately that Y is the conditional expectation of X relative to A.

Exercise 6. If $X \in L_2(\Omega, F, P)$ and A is as above, then $E[X \mid A] = \Pi X$, where Π is the orthogonal projection of $L_2(\Omega, F, P)$ onto $L_2(\Omega, A, P)$. We remind the reader that any $X \in L_2(\Omega, F, P)$ can be written (uniquely) as $X = Y + Y^{\perp}$, where $Y = \Pi X \in L_2(\Omega, A, P)$ and $Y^{\perp}$ is orthogonal to $L_2(\Omega, A, P)$ (notice that $L_2(\Omega, A, P)$ is a closed subspace of $L_2(\Omega, F, P)$).

In general, if $X \in L_1(\Omega, F, P)$, the conditional expectation $E[X \mid A]$ is the L_1-projection of X on $L_1(\Omega, A, P)$.

EXAMPLE 3. Let $X \in L_1(\Omega, F, P)$, $A_0 = \{\emptyset, \Omega\}$, and $A_1 = \{\emptyset, A, A^c, \Omega\}$, where $0 < P(A) < 1$. Then

$$E[X \mid A_0] = E[X],$$

while

$$E[X \mid A_1] = E[X \mid A]\, \mathbf{1}_A + E[X \mid A^c]\, \mathbf{1}_{A^c},$$

where

$$E[X \mid A] = \frac{E[X \mathbf{1}_A]}{P(A)},$$

namely $E[X \mid A]$ is the (elementary) conditional expectation of X given the event A. More generally, if $A_1, ..., A_n$ is a partition of Ω such that $0 < P(A_j) < 1$, $j = 1, ..., n$, and $A_n = \sigma(A_1, ..., A_n)$, then

$$E[X \mid A_n] = \sum_{j=1}^{n} E[X \mid A_j]\, \mathbf{1}_{A_j}.$$

Exercise 7. Let X, Y be r.v.'s possessing a joint p.d.f. $f(x, y)$. In elementary Probability Theory we had defined the conditional (probability) density $f(y \mid x)$ of Y, given $X = x$, as

$$f(y \mid x) = \frac{f(x, y)}{f_X(x)},$$

where

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy$$

is the (marginal) p.d.f. of X. Using $f(y \mid x)$ one can define the (elementary) conditional expectation of Y, given $X = x$, as

$$E[Y \mid X = x] = \int_{-\infty}^{\infty} y f(y \mid x) dy.$$

Show that

$$E[Y \mid X] = g(X),$$

where $g(x) = E[Y \mid X = x]$.

3.1. **Basic Properties of the Conditional Expectation.** We give a list of properties of the conditional expectation. Sometimes we will omit the obvious "a.s.'s".

Theorem 8. Let X and Y be r.v.'s, such that X and YX are integrable. If $Y \in A$, then

$$E[YX \mid A] = YE[X \mid A] \quad \text{a.s.}$$

For the proof see [1]. If $X = 1$ a.s., then $E[X \mid A] = E[1 \mid A] = 1$ a.s. (why?) and the theorem implies that

$$E[Y \mid A] = Y, \qquad \text{whenever} \quad Y \in A.$$

Some other properties:

P1 (Monotonicity)**.** If $X \leq Y$ a.s., then $E[X \mid A] \leq E[Y \mid A]$, provided $E[X]$ and $E[Y]$ exist.

P2 (Linearity)**.** If X, Y are integrable r.v.'s and $a, b \in \mathbb{R}$, then

$$E[aX + bY \mid A] = aE[X \mid A] + bE[Y \mid A].$$

P3. If X and A are independent, then $E[X \mid A] = E[X\,]$.

Exercise 8 (A Cauchy-Schwartz-type inequality)**.** If $X, Y \in L_2(\Omega, F, P)$, show that

$$E\left[|XY| \mid A\right]^2 \leq E\left[X^2 \mid A\right] E\left[Y^2 \mid A\right]$$

(Caution!).

Theorem 9 (Jensen's inequality)**.** If $\varphi : \mathbb{R} \to \mathbb{R}$ is a convex function, and X and $\varphi(X)$ are integrable r.v.'s, then

$$\varphi\left(E\left[X \mid A\right]\right) \leq E\left[\varphi(X) \mid A\right].$$

For the proof see [1].

References

[1] Kai-Lai Chung, "A Course in Probability Theory," Academic Press.
[2] H. L. Royden, "Real Analysis," Macmillan.