

Lecture 7. Expectations.

In the elementary course of probability theory, we define the expectation separately for discrete and for continuous random variables. Let me remind you these definitions.

The expectation is denoted usually with the letter E written in front of the notation of the random variable; but in some books the notation is with rounded parentheses: $E(\)$, in some other, with brackets: $E[\]$. I am going to use the notation without any kind of parentheses: just $E\xi$.

So: for discrete random variables it is taken, by definition,

$$E\xi = \sum_x x \cdot p_\xi(x), \quad (7.1)$$

where $p_\xi(x)$ is the probability mass function: $p_\xi(x) = P\{\xi = x\}$; and if the sum is infinite, we require that this infinite series should converge absolutely: $\sum_x |x| \cdot p_\xi(x) < \infty$ (absolute convergence ensures independence of the sum from the order of summation).

For a continuous random variable ξ with density $p_\xi(x)$ the definition is like this:

$$E\xi = \int_{-\infty}^{\infty} x \cdot p_\xi(x) dx; \quad (7.2)$$

just as in the discrete case, we require that the integral should converge absolutely (but for improper Riemann integrals, this turns out to be the same as requiring just that integral converge).

It turns out that these two definitions are not very convenient to base our theory on them. Many questions arise which we cannot answer. One of the questions: do random variables that are not discrete and not continuous have expectations? If they do, how to find them? Next: the very important formula

$$E(\xi_1 + \xi_2) = E\xi_1 + E\xi_2 \quad (7.3)$$

is, in some books, even not mentioned; or, at most, is mentioned and “proved”, say, for the case that ξ_1 and ξ_2 are discrete random variables (or even only for *independent* discrete random variables), when, in fact, almost an infinite variety of possible cases arises: the sum of random variables one of which is discrete. and the other continuous; or if both ξ_1 and ξ_2 are continuous random variables, their sum $\xi_1 + \xi_2$ may be not continuous; etc.

Also very important formulas are introduced: for a discrete random variable and a function $g(x)$

$$Eg(\xi) = \sum_x g(x) \cdot p_\xi(x), \quad (7.4)$$

for a continuous random variable ξ

$$Eg(\xi) = \int_{-\infty}^{\infty} g(x) \cdot p_\xi(x) dx. \quad (7.5)$$

These formulas have as their particular cases formulas (7.1), (7.2); but they do not follow immediately from them. Say, in the discrete case the random variable $\eta = g(\xi)$ is certainly a discrete random variable, and its expectation should be written, by definition, not by formula (7.4), but as

$$E\eta = Eg(\xi) = \sum_y y \cdot p_\eta(y) = \sum_y y \cdot P\{g(\xi) = y\}. \quad (7.6)$$

It is true that it is easy to deduce formula (7.4) from (7.6), grouping in (7.4) the summands with the same y and making use of additivity of the probability P ; but for continuous random variables, or for ξ and η being continuous while $\xi + \eta$ is not, we cannot do this.

The most natural way to introduce the expectation if we know theory of measure and integration is using the Lebesgue integral.

Let $\xi(\omega)$ be a random variable taking values in the extended real line $[-\infty, \infty]$ (with the corresponding Borel σ -algebra). The expectation of this random variable is defined by

$$E\xi = \int_{\Omega} \xi(\omega) P(d\omega) \quad \left[\text{or, in the shorter notation, } \int_{\Omega} \xi dP \right]. \quad (7.7)$$

The expectation, being the integral of a measurable function, may not exist; if it exists, it may be equal to ∞ or $-\infty$.

Under this definition, the property (7.3) becomes quite understandable, it is the well-known property of the (Lebesgue) integral: *if both integrals in the right-hand side of (7.3) make sense, and the sum $E\xi_1 + E\xi_2$ also makes sense (i. e., unless one of these expectations is equal to $+\infty$ and the other to $-\infty$), then the left-hand side of (7.3) makes sense, and the equality holds.*

But it is quite possible that the right-hand side of (7.3) makes no sense, while $E(\xi_1 + \xi_2)$ exists, and is even finite. E. g., we can take as ξ_1 a (real-valued, that is not taking values $\pm\infty$) random variable having no expectation, and $\xi_2 = -\xi_1$.

The only thing about this that we haven't spoken of is why from the fact that ξ_i are random variables (measurable functions) follows that so is $\xi_1 + \xi_2$. On this I am giving Problem **10**.

To get formulas (7.1), (7.2), (7.4), (7.5), we are going to use simple facts from the theory of measure and integration.

Theorem 8.1. *Let (X, \mathcal{X}) , (Y, \mathcal{Y}) be two measurable spaces, and f a measurable function (mapping) from the first space to the second one. Let m be a measure on the σ -algebra \mathcal{X} ; let $n = m \circ f^{-1}$ be the measure m carried over to the space (Y, \mathcal{Y}) by the mapping f : $n(C) = m(f^{-1}(C)) = m\{x: f(x) \in C\}$, $C \in \mathcal{Y}$ (see Lecture 3). Let g be a \mathcal{Y} -measurable function $Y \mapsto [-\infty, \infty]$.*

Then

$$\int_X g(f(x)) m(dx) = \int_Y g(y) n(dy) \quad (7.8)$$

(the precise meaning of this is: one of the sides of this formula makes sense if and only if the other side does; and in this case their values are equal).

The short notation for (7.8):

$$\int_X g \circ f \, dm = \int_Y g \, dn. \quad (7.9)$$

Since this theorem is probably too simple to be found in a book on theory of measure and integration, let us give its

Proof. According to the definition of the Lebesgue integral, we have to check (7.8) for simple measurable functions; then for arbitrary nonnegative measurable functions; and finally, for measurable functions taking values of different signs.

About which functions are we talking: about $g(y)$, or about $(g \circ f)(x) = g(f(x))$? – But if the function g is a simple (or a nonnegative) one, so is, automatically, also $g \circ f$.

So let g be a simple measurable function on (Y, \mathcal{Y}) ; i. e.,

$$g(y) = \sum_{i=1}^k c_i \cdot I_{B_i}(y), \quad (7.10)$$

where $c_i \in [-\infty, \infty]$, and B_i are disjoint subsets of Y belonging to \mathcal{Y} . Then

$$g(f(x)) = \sum_{i=1}^k c_i \cdot I_{B_i}(f(x)) = \sum_{i=1}^k c_i \cdot I_{f^{-1}(B_i)}(x), \quad (7.11)$$

because $f(x) \in B_i$ if and only if $x \in f^{-1}(B_i)$.

So we have:

$$\int_X g(f(x)) \, m(dx) = \sum_{i=1}^k c_i \cdot m(f^{-1}(B_i)) = \sum_{i=1}^k c_i \cdot n(B_i) = \int_Y g(y) \, n(dy) \quad (7.12)$$

(all terms in one of the sums are the same as those in the other, and if one of these sums makes sense, so does the other).

Now to nonnegative \mathcal{Y} -measurable g (for nonnegative functions both integrals make sense – perhaps they are nonnegative numbers, perhaps $+\infty$, but anyway they make sense).

If a nondecreasing sequence of simple nonnegative measurable functions converges to $g(y)$ at each point y :

$$0 \leq g_1(y) \leq g_2(y) \leq \dots \leq g_n(y) \leq \dots, \quad g_n(y) \rightarrow g(y) \quad (n \rightarrow \infty), \quad (7.13)$$

then also for every $x \in X$

$$0 \leq g_1(f(x)) \leq g_2(f(x)) \leq \dots \leq g_n(f(x)) \leq \dots, \quad g_n(f(x)) \rightarrow g(f(x)) \quad (n \rightarrow \infty); \quad (7.14)$$

so we have:

$$\int_X g(f(x)) \, m(dx) = \lim_{n \rightarrow \infty} \int_X g_n(f(x)) \, m(dx) = \lim_{n \rightarrow \infty} \int_Y g_n(y) \, n(dy) = \int_Y g(y) \, n(dy). \quad (7.15)$$

The reasoning from nonnegative functions to functions taking values of both signs is even simpler.

How is this theorem applied to random variables?

As the measurable space (X, \mathcal{X}) we take the sample space Ω with the σ -algebra \mathcal{F} of all events on it; instead of (Y, \mathcal{Y}) we consider an arbitrary measurable space (X, \mathcal{X}) ; as the measure m we take the probability P ; and ξ is a random variable taking values in (X, \mathcal{X}) . The measure $P \cdot \xi^{-1}$ is the distribution μ_ξ on the space (X, \mathcal{X}) . We obtain

Theorem 8.2. *If g is a measurable function on (X, \mathcal{X}) with values in the extended real line, then*

$$Eg(\xi) = \int_X g(x) \mu_\xi(dx) \quad (7.16)$$

(one side makes sense if and only if the other does; and if they do, they are equal to one another).

Particular cases of formula (7.16) are for number-valued random variables, where it is the integral from $-\infty$ to ∞ ; or for a random vector $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)$:

$$Eg(\xi_1, \dots, \xi_n) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} g(x_1, \dots, x_n) \mu_{\xi_1 \dots \xi_n}(dx_1 \dots dx_n). \quad (7.17)$$

If we remember that the Lebesgue integral with respect to a measure concentrated at countably many points x^1, \dots, x^m (\dots) is a sum, from (7.16) we obtain formulas (7.4), and, for X being the real line and the measurable function $g(x) = x$, (7.1) (and from (7.16) we get the corresponding formula involving the joint probability mass function of several random variables).

Example: if a discrete random variable ξ takes the values $1, -2, 3, -4, 5, -6, \dots, 2k-1, -2k, \dots$ with probabilities $\frac{1}{1 \cdot 2}, \frac{1}{2 \cdot 3}, \frac{1}{3 \cdot 4}, \frac{1}{4 \cdot 5}, \frac{1}{5 \cdot 6}, \frac{1}{6 \cdot 7}, \dots, \frac{1}{(2k-1) \cdot 2k}, \frac{1}{2k \cdot (2k+1)}, \dots$ (i. e., with probability mass function given by $p(2k-1) = \frac{1}{(2k-1) \cdot 2k}$, $p(-2k) = \frac{1}{2k \cdot (2k+1)}$, $k = 1, 2, 3, \dots$), the expectation does not exist, because in the series

$$\sum_{i=1}^{\infty} x^i \cdot p(x^i) = \sum_{i=1}^{\infty} \frac{(-1)^{i+1}}{i+1} \quad (7.18)$$

the positive and the negative terms taken separately both add up to $+\infty$:

$$\sum_{k=1}^{\infty} \frac{1}{2k} = \sum_{k=1}^{\infty} \frac{1}{2k+1} = \infty. \quad (7.19)$$

The expectation of the random variable $|\xi|$ exists (as of every nonnegative random variable) and is equal to $\sum_{i=1}^{\infty} \frac{1}{i+1} = \infty$.

To get formulas (7.4), (7.5) we need another simple result of the theory of measure and integration:

Theorem 8.3. Let m and n be two measures on (X, \mathcal{X}) , and suppose n has a density $f(x) (\geq 0)$ with respect to m :

$$n(C) = \int_C f(x) m(dx), \quad C \in \mathcal{X}. \quad (7.20)$$

Let $g(x)$ be a measurable function on (X, \mathcal{X}) taking values in the extended real line.

Then

$$\int_X g(x) n(dx) = \int_X g(x) \cdot f(x) m(dx). \quad (7.21)$$

(again: one side makes is defined if and only if the other one makes sense; and in this case they are equal to one another).

Proof. For simple functions... But here, in contrast with the proof of Theorem 8.1, if the function $g(x)$ is simple, the function $g(x) \cdot f(x)$ is not necessarily so: if the density $f(x)$ takes infinitely many values, $g(x) \cdot f(x)$ cannot take finitely many values (unless $c_i = 0$). And if the function $g(x) \cdot f(x)$ is simple, then $g(x)$, in all probability, isn't. So?

So: let $g(x)$ be a simple measurable function (i. e., (7.10) holds). Then we have, by definition:

$$\begin{aligned} \int_X g(x) n(dx) &= \sum_{i=1}^k c_i \cdot n(B_i) = \sum_{i=1}^k c_i \cdot \int_{B_i} f(x) m(dx) = \sum_{i=1}^k \int_X I_{B_i}(x) \cdot c_i \cdot f(x) m(dx) \\ &= \int_X \sum_{i=1}^k c_i \cdot I_{B_i}(x) \cdot f(x) m(dx) = \int_X g(x) \cdot f(x) m(dx). \end{aligned} \quad (7.22)$$

Now let $\{g_i(x)\}$ be a nondecreasing sequence of nonnegative simple functions such that $\lim_{n \rightarrow \infty} g_n(x) = g(x)$. Then also

$$0 \leq g_1(x) \cdot f(x) \leq g_2(x) \cdot f(x) \leq \dots \leq g_n(x) \cdot f(x) \leq \dots, \quad \lim_{n \rightarrow \infty} g_n(x) \cdot f(x) = g(x) \cdot f(x), \quad (7.23)$$

and

$$\int_X g(x) n(dx) = \lim_{n \rightarrow \infty} \int_X g_n(x) n(dx) = \lim_{n \rightarrow \infty} \int_X g_n(x) \cdot f(x) m(dx) = \int_X g(x) \cdot f(x) m(dx) \quad (7.24)$$

(the first equality, by definition (4.19); the last, by (4.38)).

Transition to functions $g(x)$ taking values of both signs is easy again.

Formula (7.21) is the reason for the notation $f(x) = \frac{n(dx)}{m(dx)}$ [or $\frac{dn}{dm}(x)$] for the density: we can quite formally replace $n(dx)$ by $f(x) m(dx)$ and get the correct equality.

Taking X being a Borel subset of \mathbb{R}^n with the corresponding Borel σ -algebra as \mathcal{X} , $m = \lambda_n$, $n = \mu_\xi$, and $f(x) = \frac{d\mu_\xi}{d\lambda_n}(x) = \frac{\mu_\xi(dx)}{dx} = p_\xi(x)$, we get from (7.16):

$$Eg(\xi) = \int_X g(x) \cdot p_\xi(x) dx, \quad (7.25)$$

and its particular cases (7.5) and (7.2).

In particular, if ξ has the normal distribution with parameters (a, b) , i. e., the continuous distribution with density $p(x) = \frac{1}{\sqrt{2\pi b}} e^{-(x-a)^2/2b}$, its expectation is

$$\begin{aligned} E\xi &= \int_{-\infty}^{\infty} x \cdot \frac{1}{\sqrt{2\pi b}} e^{-(x-a)^2/2b} dx \\ &= \int_{-\infty}^{\infty} a \cdot \frac{1}{\sqrt{2\pi b}} e^{-(x-a)^2/2b} dx + \int_{-\infty}^{\infty} (x-a) \cdot \frac{1}{\sqrt{2\pi b}} e^{-(x-a)^2/2b} dx \\ &= a \cdot 1 - \sqrt{b/2\pi} e^{-(x-a)^2/2b} \Big|_{-\infty}^{\infty} = a; \end{aligned} \quad (7.26)$$

if η has the exponential distribution with parameter $a (> 0)$, i. e., with density

$$p(x) = \begin{cases} 0, & x \leq 0, \\ ae^{-ax}, & x > 0, \end{cases} \quad (7.27)$$

we have:

$$Ee^{c\xi} = \int_0^{\infty} e^{cx} \cdot ae^{-ax} dx = \begin{cases} \frac{a}{a-c}, & c < a, \\ \infty, & c \geq a; \end{cases} \quad (7.28)$$

if ξ has the standard Cauchy distribution with density

$$p(x) = \frac{\pi^{-1}}{1+x^2}, \quad -\infty < x < \infty, \quad (7.29)$$

the expectation $E\xi$, finite or infinite, does not exist, because both the positive and the negative part of the integral

$$\int_{-\infty}^{\infty} x \cdot \frac{\pi^{-1}}{1+x^2} dx \quad (7.30)$$

are equal to $+\infty$.

Let me show several examples of using the fact that the expectation of the sum is equal to the sum of expectations: a basic and tremendously useful fact. It so happens that all these examples are about indicator random variables, i. e. of random variables defined by $I_A(\omega) = 1$ for $\omega \in A$, $= 0$ for $\omega \notin A$, where A is some event. The expectation of such a random variable is equal to the probability of the corresponding event: $EI_A = 1 \cdot P(A) + 0 \cdot P(A^c) = P(A)$.

Let A_1, \dots, A_n be n events. If they are mutually exclusive (disjoint), we have: $P(A_1 \cup \dots \cup A_n) = P(A_1) + \dots + P(A_n)$. Can we write a formula for $P(A_1 \cup \dots \cup A_n)$ for the case that A_i are not mutually exclusive?

We have:

$$P(A_1 \cup \dots \cup A_n) = EI_{A_1 \cup \dots \cup A_n}. \quad (7.31)$$

It is easy to see that the indicator of the intersection $I_{B_1 \cap B_2} = I_{B_1} \cdot I_{B_2}$, and the same is true for more than two B_i 's. Let us use the fact that the complement of the union is equal to the intersection of the complements:

$$\begin{aligned}
I_{A_1 \cup \dots \cup A_n} &= 1 - I_{(A_1 \cup \dots \cup A_n)^c} = 1 - I_{A_1^c} \cdot \dots \cdot I_{A_n^c} = 1 - (1 - I_{A_1}) \cdot \dots \cdot (1 - I_{A_n}) \\
&= 1 - 1 + \sum_{i=1}^n I_{A_i} - \sum_{1 \leq i < j \leq n} I_{A_i} \cdot I_{A_j} + \sum_{1 \leq i < j < k \leq n} I_{A_i} \cdot I_{A_j} \cdot I_{A_k} - \dots \\
&= \sum_{i=1}^n I_{A_i} - \sum_{1 \leq i < j \leq n} I_{A_i \cap A_j} + \sum_{1 \leq i < j < k \leq n} I_{A_i \cap A_j \cap A_k} - \dots;
\end{aligned} \tag{7.32}$$

so

$$P(A_1 \cup \dots \cup A_n) = \sum_{i=1}^n P(A_i) - \sum_{1 \leq i < j \leq n} P(A_i \cap A_j) + \sum_{1 \leq i < j < k \leq n} P(A_i \cap A_j \cap A_k) - \dots \tag{7.33}$$

Another example. Let us consider random permutations of numbers from 1 to n : the sample space Ω consists of all $n!$ sequences $\omega = (x_1, x_2, \dots, x_n)$ such that $\{x_1, x_2, \dots, x_n\} = \{1, 2, \dots, n\}$; (of course, $\mathcal{F} = \mathcal{P}(\Omega)$); and the probabilities are taken so that all different orders of the natural numbers from 1 to n are equally probable:

$$P\{(x_1, x_2, \dots, x_n)\} = \frac{1}{n!}. \tag{7.34}$$

The random variable ξ is equal to the number of numbers i , $1 \leq i \leq n$, standing in their own place: for $\omega = (x_1, x_2, \dots, x_n)$,

$$\xi(\omega) = \xi(x_1, x_2, \dots, x_n) = \#\{i: x_i = i\}. \tag{7.35}$$

The discrete random variable ξ takes values $0, 1, 2, \dots, n-3, n-2, n$ ($\xi = n-1$ is impossible: if $n-1$ numbers are in their own place, the n -th has no place to be but its own). Let us find the expectation $E\xi$.

Of course we can write: $E\xi = 1 \cdot P\{\xi = 1\} + 2 \cdot P\{\xi = 2\} + 3 \cdot P\{\xi = 3\} + \dots + (n-2) \cdot P\{\xi = n-2\} + n \cdot P\{\xi = n\}$; but it is not easy to find these probabilities.

Let us note that

$$\xi = \sum_{i=1}^n I_{A_i}, \tag{7.36}$$

where A_i is the event consisting in that the number i is in its own place: $A_i = \{(\omega = (x_1, \dots, x_n): x_i = i)\}$. So we have:

$$E\xi = \sum_{i=1}^n P(A_i). \tag{7.37}$$

Let us find the probability $P(A_i)$. Probability of every event A in our sample space is equal to $\frac{\#(A)}{\#(\Omega)} = \frac{\#(A)}{n!}$. So we have to count the number of points $\omega = (x_1, \dots, x_n)$ in the set A_i ; that is, the number of ways to place the numbers $1, \dots, n$ in the places numbered from 1 to n so that the number i is in its own place.

So we have i in the i -th place; in how many ways can we fill the remaining $n - 1$ places? It is easy to see that it is $(n - 1)!$ ways. So we have:

$$P(A_i) = \frac{(n - 1)!}{n!} = \frac{1}{n}, \quad (7.38)$$

$$E\xi = n \cdot \frac{1}{n} = 1. \quad (7.39)$$

In the same way we can find *the second moment* of the random variable ξ : $E\xi^2$. We have:

$$E\xi^2 = E(I_{A_1} + \dots + I_{A_n})^2 = E\left[\sum_{i=1}^n E(I_{A_i})^2 + \sum_{1 \leq i, j \leq n, i \neq j} E(I_{A_i} \cdot I_{A_j})\right]. \quad (7.40)$$

As for $E(I_{A_i})^2$, we have $(I_{A_i})^2 = I_{A_i}$, because this random variable takes only two values, and $0^2 = 0$, $1^2 = 1$; so $E(I_{A_i})^2 = EI_{A_i} = 1/n$. For $E(I_{A_i} \cdot I_{A_j}) = EI_{A_i \cap A_j} = P(A_i \cap A_j)$, $i \neq j$, we have:

$$P(A_i \cap A_j) = \frac{\#(A_i \cap A_j)}{\#(\Omega)} = \frac{(n - 2)!}{n!} = \frac{1}{n(n - 1)} \quad (7.41)$$

(the same reasoning: we have i and j put in their own places, and have to fill the remaining $n - 2$ places with the remaining $n - 2$ numbers). There are $n(n - 1)$ summands with $i \neq j$, so

$$E\xi^2 = n \cdot \frac{1}{n} + n(n - 1) \cdot \frac{1}{n(n - 1)} = 2. \quad (7.42)$$

See Problem 18.