

Lecture 3.

The lecture notes I am writing do not follow exactly the lectures; some things in them are given in a different order, and some things not in the lectures are included.

How can we work with σ -algebras generated by class of sets? We can do so in some indirect ways. The following Microtheorem is useful here.

Theorem 3.1. *Suppose \mathcal{A} and \mathcal{C} are classes of subsets of a space X . If*

$$\mathcal{A} \subseteq \sigma(\mathcal{C}), \quad \mathcal{C} \subseteq \sigma(\mathcal{A}), \quad (3.1)$$

then

$$\sigma(\mathcal{A}) = \sigma(\mathcal{C}). \quad (3.2)$$

Proof. It is clear that if $\mathcal{D} \subseteq \mathcal{E}$, then $\sigma(\mathcal{D}) \subseteq \sigma(\mathcal{E})$. Applying this to the first inclusion in (3.1), we get:

$$\sigma(\mathcal{A}) \subseteq \sigma(\sigma(\mathcal{C})) = \sigma(\mathcal{C}) \quad (3.3)$$

(the last equality is quite clear). From the second inclusion in (3.1) we obtain, the same way, that $\sigma(\mathcal{C}) \subseteq \sigma(\mathcal{A})$; which, together with (3.3), yields (3.2).

Let us show an example of how this is used.

Let \mathcal{A} be the class of all intervals, (a, b) , $(a, b]$, $[a, b)$, $[a, b]$, finite or infinite, in the real line (a one-point set $\{a\}$ is also an interval: $\{a\} = [a, a]$: it is the set of all real x such that $a \leq x \leq a$; and the empty set can also be considered as an interval). Let \mathcal{C} be the class of all intervals, finite or infinite, of the form $(a, b]$ (we understand what the interval $(-\infty, b]$ is; but what is $(a, \infty]$? – By definition, it is the set $\{x \in \mathbb{R}^1 : a < x \leq \infty\}$; that is, the same as $\{x \in \mathbb{R}^1 : a < x < \infty\} = (a, \infty)$: no real number can be *equal* to ∞).

Clearly $\mathcal{A} \supseteq \mathcal{C}$. Let us prove that the σ -algebras $\sigma(\mathcal{A})$ and $\sigma(\mathcal{C})$ are the same.

According to Theorem 3.1, it is enough to check that $\mathcal{A} \subseteq \sigma(\mathcal{C})$. This means that every interval of any kind ($(\ , \)$, $[\ , \)$, $[\ ,]$) belongs to $\sigma(\mathcal{C})$. And to do this, it is enough to represent every interval as the result of applying countably many set-theoretic operations (\cup , \cap , and c) to intervals of the form $(\ ,]$.

An interval (a, b) is the union of smaller semi-closed intervals:

$$(a, b) = \bigcup_{i=k}^{\infty} (a, b - 1/i] \quad (3.4)$$

(make a picture; I did not write the union from **1** to infinity because I did not want questions about what happens if the length of the interval (a, b) is less than 1. But in fact, it would be OK: several first summands in (3.4) would be intervals with the “right” end $b - 1/i$ to the left of the “left” end a ; such intervals are just empty, and it wouldn’t affect the union). So (a, b) belongs $\sigma(\mathcal{C})$ by the σ -algebra axiom **3 σ** (see Lecture 1).

The interval $[a, b]$ is the complement of the union

$$(-\infty, a) \cup (b, \infty] \tag{3.5}$$

(if $a = -\infty$ or $b = \infty$, the corresponding intervals are empty), and such intervals are both in $\sigma(\mathcal{C})$; so $[a, b] \in \sigma(\mathcal{C})$. Finally, the interval $[a, b]$ is the complement of $(-\infty, a) \cup [b, \infty]$, and also belongs to $\sigma(\mathcal{C})$.

The same σ -algebra is also the same as the σ -algebra generated by all semi-infinite intervals $(-\infty, b]$, $-\infty < b < \infty$ (because $(a, b] = ((-\infty, a] \cup (-\infty, b]^c)^c$), and the same as that generated by all open subsets of \mathbb{R}^1 (the last statement is true because every open subset of the real line is a countable union of open intervals).

The σ -algebra

$$\sigma\{\text{all intervals}\} = \sigma\{(a, b]: -\infty \leq a \leq b \leq \infty\} = \sigma\{(-\infty, b]: b \in \mathbb{R}^1\} = \sigma\{\text{all open sets}\} \tag{3.6}$$

that we introduced is very important for us. Remember that I said that the choice of the sample space Ω in applications of probability theory is natural, related to the nature of the experiment whose mathematical model we are trying to build; and the choice of the σ -algebra \mathcal{F} of its subsets proclaimed as events is *standard*. Namely, if the sample space Ω is countable (possibly, finite), we take $\mathcal{F} = \mathcal{P}(\Omega)$, the class of *all* subsets of Ω ; and if Ω is the real line, or part of it, or a set in \mathbb{R}^n ..., then the choice of the σ -algebra \mathcal{F} is *also standard*.

The time has come to set this standard. If $\Omega = \mathbb{R}^1$, we take as \mathcal{F} the σ -algebra (3.6).

This σ -algebra is called the (one-dimensional) *Borel σ -algebra*, and its elements are called *Borel sets*. The notation for it that we are going to use is \mathcal{B}^1 .

When Henri Lebesgue constructed what we call now the Lebesgue measure λ_1 on the real line, it was defined on some σ -algebra \mathcal{L}^1 of subsets of \mathbb{R}^1 , called *measurable* subsets. This σ -algebra contains all intervals, and the Lebesgue measure of every interval is equal to its length. Since \mathcal{L}^1 is *some* σ -algebra containing all intervals, and \mathcal{B}^1 is *the smallest* such σ -algebra, we have clearly

$$\mathcal{L}^1 \supseteq \mathcal{B}^1.$$

Is the inclusion \supseteq , in fact, a *strict* inclusion \supset , or is $\mathcal{L}^1 = \mathcal{B}^1$? You can be sure that mathematicians have worked to ascertain this; it turned out that \mathcal{L}^1 is *wider* than \mathcal{B}^1 : $\mathcal{L}^1 \supset \mathcal{B}^1$ (and even *much* wider – though I don't want to spend time explaining what it means).

So which of these σ -algebras should we use as the standard?

It turns out that the σ -algebra \mathcal{B}^1 of Borel sets is *very* large: quite large enough for all practical purposes, and for most of theoretic ones. As a matter of fact, we cannot even *construct an example* of a non-Borel set; and we learn about existence of such only by indirect methods.

So in our probability theory we are quite satisfied with the Borel σ -algebra \mathcal{B}^1 , and we don't need in fact any sets belonging to \mathcal{L}^1 but not to \mathcal{B}^1 . We even don't need to know that \mathcal{L}^1 is strictly larger than \mathcal{B}^1 . (The question arises: why did Lebesgue introduce the σ -algebra \mathcal{L}^1 if \mathcal{B}^1 is quite enough? – But this was discovered only later). The second reason that we are using the Borel σ -algebra rather than some larger ones is that the σ -algebra \mathcal{L}^1 is closely related to a specific measure: the Lebesgue measure λ_1 ; and

for other measures on the real line the σ -algebras on which one can consider them are different from \mathcal{L}^1 . It would be unreasonable to consider many different σ -algebras on the same space \mathbb{R}^1 if we can do with just one, \mathcal{B}^1 .

We can consider Borel σ -algebras in multidimensional spaces \mathbb{R}^n ; these σ -algebras can be defined either as generated by all multidimensional “intervals” – i. e., rectangles $(a, b] \times (c, d]$, or parallelepipeds in the three-dimensional case, etc.; or as generated by all open sets. Problem **7** given to you is about the fact that it is all the same. The Borel σ -algebra in the n -dimensional Euclidean space is denoted \mathcal{B}^n .

We can also define the Borel σ -algebra \mathcal{B}_X in every space X that is a metric space, or even a non-metric topological space: in every space where a class of open sets is defined. Of course, since there needn't be any “intervals” or rectangles in the space X , the σ -algebra \mathcal{B}_X is defined as one generated by the class \mathcal{O} of open subsets of X .

I said that standard σ -algebras are used in spaces that are either Euclidean spaces \mathbb{R}^n , or *their parts*; but I have spoken only of the whole spaces.

If X is a Borel subset of \mathbb{R}^n , we can define the σ -algebra \mathcal{B}_X either as one generated by subsets of X that are open *in* X (a set $A \subseteq X$ is called open in X if for every point $x_0 \in A$ there is its neighborhood in X that is entirely within A : there exists a positive radius r such that $\{x \in X: |x - x_0| < r\} \subseteq A$); or as the σ -algebra consisting of all Borel subsets of A . See Problem **8**.

Now we go to random variables.

First, a piece belonging to the set-theoretic introduction to probability and measure theory: what a *measurable function* is.

When Lebesgue introduced the concept of measurability, it was understood as measurability *with respect to the Lebesgue measure*. But afterwards mathematicians understood that this concept can be formulated without reference to any measure: as one belonging to the set-theoretic introduction.

A pair (X, \mathcal{X}) of a space X and a σ -algebra \mathcal{X} in it is called a *measurable space*.

Let (X, \mathcal{X}) and (Y, \mathcal{Y}) be two measurable spaces. Let $f(x)$ be a function $f: X \mapsto Y$ (i. e., a function defined on X , with values in Y). We call this function $(\mathcal{X}, \mathcal{Y})$ -*measurable* (or measurable with respect to \mathcal{X}, \mathcal{Y}) if for every set $C \in \mathcal{Y}$ its inverse image $f^{-1}(C) = \{x: f(x) \in C\}$ belongs to \mathcal{X} :

$$C \in \mathcal{Y} \Rightarrow f^{-1}(C) \in \mathcal{X}. \tag{3.7}$$

If we consider a number-valued or a vector-valued function ($Y = \mathbb{R}^1$ or \mathbb{R}^n), and take the standard σ -algebra \mathcal{B}^1 or \mathcal{B}^n as the σ -algebra \mathcal{Y} , we omit the mention of \mathcal{Y} and say “ \mathcal{X} -measurable”, or “measurable with respect to the σ -algebra \mathcal{X} ”. If \mathcal{X} is a Borel σ -algebra, we call a function f that is measurable with respect to it *Borel measurable*.

Let Ω be a sample space, \mathcal{F} , a σ -algebra in it whose elements we proclaim events. Let (X, \mathcal{X}) be a measurable space. A random variable taking values in this measurable space is, by definition, a measurable function from (Ω, \mathcal{F}) to (X, \mathcal{X}) . We will denote random variables with greek letters. So, a random variable with values in (X, \mathcal{X}) is a

function $\xi: \Omega \mapsto X$ that is $(\mathcal{F}, \mathcal{X})$ -measurable:

$$C \in \mathcal{X} \Rightarrow \xi^{-1}(C) = \{\omega: \xi(\omega) \in C\} \text{ [the short notation: } \{\xi \in C\} \in \mathcal{F} \text{ (is an event)}]. \quad (3.8)$$

Of course, for the most part we consider $X = \mathbb{R}^1$ (just *random variables*, without any mention of the space \mathbb{R}^1 in which they take values); or $X = \mathbb{R}^n$ (random *vectors*); or, say, the space of all $n \times n$ matrices (or *symmetric* matrices) – then we speak of *random matrices* (or random symmetric matrices). In all these cases as the σ -algebra \mathcal{X} we take the corresponding Borel σ -algebra.

Theorem 3.2. *Let f be a measurable function from (X, \mathcal{X}) to (Y, \mathcal{Y}) , and g a measurable function from (Y, \mathcal{Y}) to (Z, \mathcal{Z}) . Then the composition $g \circ f$ (i. e. the function $X \mapsto Z$ defined by $(g \circ f)(x) = g(f(x))$) is an $(\mathcal{X}, \mathcal{Z})$ -measurable function.*

The **proof** is so simple that I omit it.

The “probabilistic” formulation (i. e., in the language of (the set-theoretic introduction to) probability theory):

Let ξ be a random variable with values in (X, \mathcal{X}) . If g is an $(\mathcal{X}, \mathcal{Z})$ -measurable function $X \mapsto Z$, then $\zeta = g(\xi)$ (wich is the short notation for $\eta(\omega) = g(\xi(\omega))$) is a random variable with values in (Z, \mathcal{Z}) .

The particular case of $X = Z = \mathbb{R}^1$, $\mathcal{X} = \mathcal{Z} = \mathcal{B}^1$:

Let ξ be a random variable (by default, taking values in the real line), and g a Borel measurable real-valued function. Then $g(\xi)$ also is a random variable.

The definition (3.7) requires checking $f^{-1}(C) \in \mathcal{X}$ for *all* sets in \mathcal{Y} . This may be too many sets.

But usually we are able to do it for much fewer sets:

Theorem 3.3. *Let \mathcal{Y} be the σ -algebra generated by some class \mathcal{A} of subsets of Y . Then if*

$$C \in \mathcal{A} \Rightarrow f^{-1}(C) \in \mathcal{X}, \quad (3.9)$$

the function f is $(\mathcal{X}, \mathcal{Y})$ -measurable.

Proof. Let us denote with \mathcal{D} the class of all subsets of Y for which $f^{-1}(C) \in \mathcal{X}$.

Clearly $\mathcal{D} \supseteq \mathcal{A}$.

If we prove that \mathcal{D} is a σ -algebra, then, by definition, $\mathcal{D} \supseteq \sigma(\mathcal{A}) = \mathcal{Y}$, and $\mathcal{Y} \subseteq \mathcal{D}$ is just the statement of our theorem.

So let us check that \mathcal{D} is a σ -algebra in Y .

This means, first, that

$$Y \in \mathcal{D}; \quad (3.10)$$

second and third, that

$$C \in \mathcal{D} \Rightarrow C^c = Y \setminus C \in \mathcal{D}, \quad (3.11)$$

$$C_1, C_2, \dots, C_n, \dots \in \mathcal{D} \Rightarrow \bigcup_{i=1}^{\infty} C_i \in \mathcal{D}. \quad (3.12)$$

Checking (3.10): $f^{-1}(Y) = \{x: f(x) \in Y\} = X \in \mathcal{X}$. Now to (3.11):

$$f^{-1}(Y \setminus C) = X \setminus f^{-1}(C) \in \mathcal{X}; \quad (3.13)$$

and (3.12):

$$f^{-1}\left(\bigcup_{i=1}^{\infty} C_i\right) = \bigcup_{i=1}^{\infty} f^{-1}(C_i) \in \mathcal{X}. \quad (3.14)$$

So, e. g., since the one-dimensional Borel σ -algebra is generated by all semi-infinite intervals $(-\infty, a]$ (or by all $(-\infty, a)$, $-\infty < a < \infty$), to check that a real-valued function $\xi = \xi(\omega)$ is a random variable it is enough to check that all sets $\{\omega: \xi(\omega) \leq a\}$ (or $<$) are *events*.

The proof of Theorem 3.3 is the standard thing that we have to do infinitely many times if we want to set probability theory rigorously; we cannot do without such things, but they are pretty simple, and we have to remember that they are not what is very important in probability theory.

Theorem 3.4. *Every continuous function $f: X \mapsto Y$ (of course, if we want to consider continuous functions, we have to suppose that X and Y are sets in Euclidean spaces, or metric, or at least topological spaces) is $(\mathcal{B}_X, \mathcal{B}_Y)$ -measurable (in short: Borel measurable).*

Proof. By definition, \mathcal{B}_Y is the σ -algebra generated by the open subsets of Y ; so by Theorem 3.3 it is enough to check that for every open $C \subseteq Y$

$$f^{-1}(C) = \{x: f(x) \in C\} \in \mathcal{B}_X \quad [= \sigma\{\text{open subsets of } X\}]. \quad (3.15)$$

But if the function f is continuous, the inverse image of every open set is again open; so (3.15) is true.

In the same way we prove that *every monotone function $f: \mathbb{R}^1 \mapsto \mathbb{R}^1$ is Borel measurable*: we use the fact that $\mathcal{B}^1 = \sigma\{\text{all intervals}\}$, and that the inverse image f^{-1} of every interval is again an interval.

Now, the most important thing in probability theory must be the *probabilities*; and we haven't even mentioned them. Of course it is because we just defined what random variables are, and this is only the first, and not the most important thing about random variables.

One of the central things about random variables is their *distributions*.

There are several different terminologies in probability theory. In some of them, the word "distribution" is used as a *term*: there is a certain class of mathematical objects that are called *distributions*. In some other terminologies, this word is used as a keyword: the common part of some terms used for some things that are in some way similar to one another: *distribution function, distribution density, etc.*

In these lectures, I am going to use the word "distribution" as a term.

Now we go a little to the general measure theory.

I would like to tell you that measure theory: the theory of measure and integration, contains material of different degrees of difficulty: some things are very simple; some other are of medium difficulty (such is,

for example, the construction of Lebesgue integral); and some other parts are pretty complicated. Such is the construction of the Lebesgue measure on the real line or in \mathbb{R}^n . We could develop our theory using only very simple and medium-simple things; but in probability theory this would restrict us to discrete random variables, and we wouldn't be able to speak of continuous distributions, for which integration with respect to the Lebesgue measure is needed. So in fact we cannot do without the Lebesgue measure, and with it, the more complicated things in measure theory.

What follows belongs to the *very simple* part of measure theory.

Let $\xi: \Omega \rightarrow X$ be a random variable taking values in a measurable space (X, \mathcal{X}) . The *distribution* of ξ is defined as a set function $\mu(C) = \mu_\xi(C)$ on the σ -algebra \mathcal{X} defined by

$$\mu_\xi(C) = P\{\xi \in C\}, \quad C \in \mathcal{X}. \quad (3.16)$$

For example, if ξ is just a random variable (i. e., real-valued one), μ_ξ is defined on \mathcal{B}^1 ; if ξ is an n -dimensional random vector, μ_ξ is defined on \mathcal{B}^n ; etc.

It turns out that the distribution of a random vector is necessarily *a measure*.

In fact, this is not specific to probability theory: this is a general thing in measure theory.

Suppose f is a measurable function from a measurable space (X, \mathcal{X}) into another measurable space (Y, \mathcal{Y}) . If m is a measure on \mathcal{X} , it is automatically carried to the σ -algebra \mathcal{Y} , creating there another measure – denote it by n .

Namely, we take, for $C \in \mathcal{Y}$,

$$n(C) = m(f^{-1}(C)). \quad (3.17)$$

A reasonable notation for the new measure n is $n = m \circ f^{-1}$; but we still have to prove that this is a measure.

Proof. We have to check that $n(C)$ is nonnegative (but this is clear); that

$$n(\emptyset) = 0; \quad (3.18)$$

and that for $C_1, C_2, \dots, C_n, \dots \in \mathcal{Y}$

$$C_i \text{ disjoint} \Rightarrow n\left(\bigcup_{i=1}^{\infty} C_i\right) = \sum_{i=1}^{\infty} n(C_i). \quad (3.19)$$

About the empty set it is clear, because $f^{-1}(\emptyset) = \emptyset$. The countable additivity of n :

$$\begin{aligned} (m \circ f^{-1})\left(\bigcup_{i=1}^{\infty} C_i\right) &= m\left(f^{-1}\left(\bigcup_{i=1}^{\infty} C_i\right)\right) = m\left(\bigcup_{i=1}^{\infty} f^{-1}(C_i)\right) \\ &= \sum_{i=1}^{\infty} m(f^{-1}(C_i)) = \sum_{i=1}^{\infty} (m \circ f^{-1})(C_i) \end{aligned} \quad (3.20)$$

follows from countable additivity of m .

So the distribution μ_ξ of a random variable ξ is a measure; and it may be denoted as $P \circ \xi^{-1}$.

Clearly, the distribution of every random variable is such that its value at the largest set (X if this is a random variable taking values in (X, \mathcal{X})) is equal to 1. We will call such measures *probability measures*.

There are different sorts of random variables, and concrete forms of their distributions.

A random variable ξ is called a *discrete* random variable if it takes a countable number of values: x^1, x^2, \dots, x^m (\dots) (finite, or countably infinite).

If we know what is usually called the *probability mass function*

$$p(x) = p_\xi(x) = P\{\xi = x\} \quad [= \mu_\xi\{x\}] \quad (3.21)$$

of a discrete random variable ξ , we can find its distribution. Namely, since

$$\{\xi \in C\} = \bigcup_{i: x^i \in C} \{\xi = x^i\}, \quad (3.22)$$

where the events-summands are disjoint, we have by countable additivity:

$$\mu_\xi(C) = P\{\xi \in C\} = \sum_{i: x^i \in C} P\{\xi = x^i\} = \sum_{i: x^i \in C} p(x^i) \quad (3.23)$$

(or just $\mu_\xi(C) = \sum_{x \in C} p_\xi(x)$ – taking that in a sum, even if it is an uncountable one, zeros don't count).

Of course the measure μ_ξ for a discrete random variable is defined in the same way for *all* subsets of X (for all sets in $\mathcal{P}(X)$), and (3.23) holds for arbitrary $C \in \mathcal{P}(X)$; but we don't need this: say, the Borel σ -algebra is quite large enough in the case of X being a subset of \mathbb{R}^n .

The probability mass function of every discrete random variable, obviously, satisfies the following:

$$p(x) \geq 0, \quad \sum_x p(x) = 1 \quad (3.24)$$

(a sum over x without mentioning the range within which this x is changing denotes the sum over *all* possible values of x ; that this sum is equal to 1 follows from the fact that it is equal to $P\{\xi \in X\}$).

If $p(x)$ is a function satisfying (3.24), there exists a probability space (Ω, \mathcal{F}, P) and a random variable ξ on it such that $p_\xi(x) = P\{\xi = x\}$ is the given function $p(x)$.

Indeed, we take $\Omega = \{x: p(x) > 0\}$ – this clearly is a countable (finite or infinite) set; $\mathcal{F} = \mathcal{P}(\Omega)$, the class of all its subsets; and define the probability P by

$$P(C) = \sum_{x \in C} p(x). \quad (3.25)$$

Clearly (again) this P satisfies the axioms. The random variable ξ is defined by

$$\xi(x) = x, \quad x \in \Omega. \quad (3.26)$$

It is easy to see that P works at the same time as the distribution μ_ξ of our random variable, and $P\{\xi = x\}$ is the sum consisting of exactly one summand, namely $p(x)$.

So we can call distributions that are given by formula (3.25) (i. e., $\mu(C) = \sum_{x \in C} p(x)$) with $p(x)$ satisfying (3.24) *discreet distributions*; but we should keep in mind that a random variable having a discrete distribution may be not discrete: it may take uncountably many values different from x^1 , x^2 , ..., x_n (, ...), but all of them combined with zero probability.

Now, the other important class of random variables, or rather that of distributions, is that of (*absolutely*) *continuous* distributions; or distributions with *densities*.

We say that the distribution μ_ξ of a random variable ξ taking values in $X \subseteq \mathbb{R}^n$ is (*absolutely*) *continuous* if there exists a nonnegative function $p(x)$ on X such that

$$\mu_\xi(C) = P\{\xi \in C\} = \int_C p(x) dx \quad (3.27)$$

(in the multidimensional case x can be written as (x_1, \dots, x_n) , and dx means $dx_1 \dots dx_n$).

Again, the concept of *density* is used not only in probability theory, but also in the general measure theory.

Let (X, \mathcal{X}) be a measurable space; m , a measure on it. Suppose n is another set function defined on \mathcal{X} . We say that n has a density $f(x)$ with respect to m if

$$n(C) = \int_C f(x) m(dx). \quad (3.27)$$

The integral here is (of course) a Lebesgue integral. But reminding what Lebesgue integral with respect to an arbitrary measure is, and explaining how to formulate all this more accurately, is left till the next lecture.