

Lecture note 1. Probability theory.

The lecture notes are not identical with the lectures; sometimes the material will be organized differently. In particular, in this lecture note are included some things touched upon in Lecture 2.

Stochastic equations are an important tool in the theory of stochastic processes; which, in its turn, is an important field in probability theory. So we start with probability.

It was discovered in the 1930's that the natural base on which one should build probability theory is *measure theory*. I do not assume that the students in this course have taken or are taking now a course in measure theory; so why mention this at all? I will be using *the language* of measure theory; and this is quite a different thing. Measure theory – the theory of measure and integration – is a profound theory, and some of its results are difficult; but its *language* is useful and simple. The situation is the same as with the set theory: this theory also has some profound and difficult results; but we use, without referring to this theory properly speaking, its language very widely: we speak constantly about *sets*, about mappings (functions) from one set to another, etc.

Also I will be giving some probabilistic results without proof, their proof being based on measure theory: in such cases I am going to refer to measure theory.

And if this course will result in some student deciding to take a course in theory of measure and integration, I won't regret it.

So to probability theory.

In the background of every problem of probability theory lies a *probability space*, which is a triple of mathematical objects:

- The *sample space*, which we are going to denote with the letter Ω (the capital Greek “omega”); its generic element, a *sample point*, will be denoted ω (the small Greek “omega”: *not* the letter w). In the elementary probability course, the notations usually are: S for the sample space, and s for a generic sample point. We are using different notations. The reasons are, first, the tradition – and second, we are going to consider stochastic *processes*, that develop in *time*; a natural notation for the time variable is t – and if we want to consider a second time variable, it will be denoted with s .

- The class \mathcal{F} of subsets of Ω that are called *events*.

What is a *class of sets*? This means just some *set of sets*: a set consisting of sets; but if we say “set of sets”, our tongue may get twisted, so we use the word *class* instead. So: class of sets, class of subsets.

In the elementary course of probability theory we usually say just that an *event* is a subset $A \subseteq \Omega$ (the sign \subseteq means that the subset to the left of this sign is a part of that to the right, including the possibility that this part coincides with the whole set), disregarding the fact that possibly not every subset of Ω is an event. I'll speak about why a little later.

- The probability P , which is a function $P : \mathcal{F} \mapsto \mathbb{R}$: a real-valued function defined on *events*. The value $P(A)$ of this function at an event A is called *the probability of the event A* .

These three objects should satisfy the following axioms:

There are practically no axioms about the sample space Ω : we can take as Ω an arbitrary set, consisting of objects ω of an arbitrary nature; the only requirement is that this set shouldn't be empty.

The axioms for the class \mathcal{F} are:

- The set Ω itself belongs to \mathcal{F} (remember that Ω is a subset of Ω); that is, we should be able to consider the event *that is sure to occur*: that contains *every* sample point ω .

- If $A \in \mathcal{F}$ (i. e., A is an event), then its complement $A^c = \Omega \setminus A \subseteq \mathcal{F}$ (i. e., we should be able to consider the *opposite event*).

- If $A_1, A_2, \dots, A_n, \dots$ is a sequence of events (i. e., $A_i \in \mathcal{F}$ for $i = 1, 2, \dots, n, \dots$), then their union $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$ (that is, we can consider the event consisting in that at least one of the events A_i occurs).

The axioms of the probability P are, in fact, those of *measure*; so here I introduce the first term of the language of measure theory.

Let X be an arbitrary set, and \mathcal{X} some class of its subsets. Suppose a nonnegative function m is defined on the class \mathcal{X} : a function $m: \mathcal{X} \mapsto [0, \infty]$ (we allow the value ∞ for the function m). The set function m is called a *measure* if the following axiom is satisfied:

- If $A_1, A_2, \dots, A_n, \dots$ is a sequence of disjoint sets belonging to \mathcal{X} (*disjoint* means that their intersections are empty: $A_i \cap A_j = \emptyset$ for $i \neq j$), and $\bigcup_{i=1}^{\infty} A_i \in \mathcal{X}$, then

$$m\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} m(A_i). \quad (1.1)$$

This property is called that of *countable additivity*.

Examples of measures.

- X is the plane \mathbb{R}^2 . There exists a class $\mathcal{X} = \mathcal{L}^2$ of its subsets (called the class of *Lebesgue measurable* planar sets) and a measure λ_2 on it (called the 2-dimensional Lebesgue measure) such that for every rectangle $[a, b] \times [c, d] = \{(x, y) : x \in [a, b], y \in [c, d]\}$ we have: $\lambda_2([a, b] \times [c, d]) = (b - a) \cdot (d - c)$. In other words, the usual area in the plane is a measure.

As for the *final* additivity of area, it was known as far back as to Euclid: representing polygons as unions of triangles, and putting triangles together to get a rectangle was the way in which areas of figures were calculated. *Countable* additivity (1.1) was known, in fact, in some particular cases, to Archimedes. The modern theory of areas (and volumes, and lengths), with the class \mathcal{L}^2 so vast that one cannot really construct a set $A \subset \mathbb{R}^2$ not belonging to \mathcal{L}^2 (but one can give a *proof*, even if it is an ineffective one, of the fact that such planar sets exist) was developed at the end of the nineteenth century and the beginning of the twentieth by French mathematicians Henri Lebesgue and Emil Borel.

Note that there are sets A whose area is equal to ∞ : for example, $\lambda_2(\mathbb{R}^2) = \infty$.

- Let X be an arbitrary non-empty set, and let \mathcal{X} be the class of *all* its subsets. Let x_0 be a point in X . Define the function $m: \mathcal{X} \mapsto [0, \infty]$ by

$$m(A) = \begin{cases} 1 & \text{if } A \ni x_0, \\ 0 & \text{if } A \not\ni x_0. \end{cases} \quad (1.2)$$

The function m is a measure.

Indeed, for a sequence of disjoint sets either no set of this sequence contains the point x_0 , and the equality (1.2) is satisfied because it is zero being equal to the sum of a series whose all terms are zero; and if one of A_i does contain x_0 , then it is *only* one of them (another A_j cannot contain x_0 because it would mean that $A_i \cap A_j \neq \emptyset$), and the equality (1.2) is that 1 is equal to the sum of a series with one term being 1 and the rest equal to 0.

A third particular case of a measure is *probability*. So let us go to probabilities.

The axioms for the probability P are:

- P is a measure on the class \mathcal{F} of all events.
- $P(\Omega) = 1$.

Deciphering these axioms, we get:

- For every event A we have $0 \leq P(A) \leq 1$.
- If $A_1, A_2, \dots, A_n, \dots$ is a sequence of disjoint events (which means: no two events of this sequence can occur together), then $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

Since you are not supposed to know measure theory, we won't get much profit of the fact that the probability is a measure; but it is useful to keep it in mind.

As for the axioms concerning the class \mathcal{F} , axioms of exactly the same kind are considered also in measure theory. Both in measure theory and in probability theory there is a class of problems about this or that set belonging to a specified class of sets (\mathcal{F} , in probability theory). It is necessary to solve problems of this kind if we want to build our theory with absolutely no gaps. These problems do not belong to measure theory or to probability theory properly speaking: they belong rather to a *set-theoretic introduction* that is common to both these theories. The methods of solving these problems are not very complicated, and they are rather standard. I am going to disregard these problems – so probability theory and the theory of stochastic processes in our course will be built with some gaps; but I do not want to spend time on such technical questions when there is so much more interesting material to speak about.

Why introduce the class \mathcal{F} of events at all: isn't it possible to spend some more effort and define probability P for *all* subsets of Ω , as we claim it done in the elementary course of probability theory? It turns out (not a very simple result in measure theory) that this is impossible:

One cannot construct a probability measure on all subsets of an interval, say, $(0, 1]$.

This statement is very good emotionally, showing the tragic situation in measure theory and probability theory: they *cannot* be simplified this way; but unfortunately, it is false: formula (1.2) does define a measure on *all* subsets of the set X , which can be chosen to be an interval. The precise formulation of the fact from measure theory I am referring to is a little more complicated:

One cannot construct a probability measure on all subsets of an interval, say, $(0, 1]$, such that the measure of every one-point set is equal to 0.

That is, the only cases when we can define probabilities for all subsets of the sample space are those similar to (1.2) or a little more complicated: measures concentrated at a countable number of points. This would exclude at once a big part of probability theory: the part where random variables with continuous distributions are considered.

But about random variables we'll speak a little later.

In the elementary probability course, we speak not only about the theory, but also about its possible applications. Some *principles* of these applications are formulated. I'll speak now about one such principle:

If the probability $P(A)$ is small enough, we can disregard the possibility that the event A will occur.

This principle is the basis of very many applications of probability theory, in particular, for problems of mathematical statistics having to do with constructing confidence intervals or with testing statistical hypotheses.

The principle I formulated above is not formulated in the frank an honest, unambiguous mathematical way: what does it mean “small enough”? If it said: “if $P(A) < 0.05$ ”, or “if $P(A) \leq 0.0001$ ” – then it would be mathematically understandable.

But of course, the place for precise mathematical statements is within mathematics; the principle in question is just outside mathematics: at the borderline between mathematics and the extra-mathematical world (in which the potential applications of probability theory would take place), and absolute precision would jeopardize the applicability of the principle. *How small* the probability $P(A)$ should be in order for us to be able to disregard the possibility of A is a *practical* question, not one of the theory.

But there is exactly one value of the probability that should be considered as small enough for *all* practical purposes: namely, 0. So if $P(A) = 0$, we can disregard the possibility of the event A occurring.

And in probability theory we systematically disregard events of zero probability.

By the way, this is a common trait of probability theory and measure theory. It goes as far back as to ancient Greeks: considering areas, they disregarded *lines* having zero area; and they even did not think about whether to include the border of a triangle in it, or not.

Having this in mind, I am going to introduce the following term of probability theory: a property of a sample point ω is said to be satisfied *almost surely* if it is satisfied except for a set of ω 's having zero probability. For example: if $\Omega = (0, 1]$, and the probability P is the length (in the absolutely precise mathematical language: the one-dimensional Lebesgue measure λ_1 , defined on some class $\mathcal{F} = \mathcal{L}^1$ of subsets of this interval), the property $\omega < 1$ is not satisfied for all $\omega \in \Omega$, but it is satisfied *almost surely*: with the exception of $\omega = 1$ – and the length of the one-point exceptional set $\{1\}$ is definitely equal to 0.

The next concept of probability theory: a *random variable*; and the *distribution* of a random variable.

A random variable is a function $X : \Omega \mapsto \mathbb{R}$ – a real-valued function $X = X(\omega)$ defined on the sample space – that has the following property:

- For every interval $I \subseteq \mathbb{R}$ the set $\{\omega : X(\omega) \in I\}$ is an event: $\{\omega : X(\omega) \in I\} \in \mathcal{F}$.

Two random variables X and Y (on the same probability space (Ω, \mathcal{F}, P)) are called equivalent: $X \sim Y$, if they are equal to one another almost surely:

$$P\{X \neq Y\} = 0 \tag{1.3}$$

(or, which is the same, $P\{X = Y\} = 1$). We are going to systematically disregard the distinction between equivalent random variables.

The word *distribution* is used in probability theory in two different ways. Either it is used as some sort of key word to speak of several similar concepts: e. g., if we are told to find the distribution of a random variable, it means finding what is called “probability mass function” if this is a discrete random variable, or finding the distribution density if it is a continuous random variable, or finding the cumulative distribution function. Or the word “distribution” may be used as a precise mathematical term. In this course, we are choosing the second way.

The *distribution of the random variable* X is, by definition, a function $\mu = \mu_X$ of subsets of the real line \mathbb{R} defined by

$$\mu(C) = \mu_X(C) = P\{X \in C\}. \quad (1.4)$$

The notation $P\{X \in C\}$ is short for $P(\{\omega : X(\omega) \in C\})$. You see: something is standing under the sign of probability. What can it be? Probabilities can be only of *events*, and every event is a set consisting of sample points ω . So mentioning ω in the notation $P(\{\omega : X(\omega) \in C\})$ is redundant. Also: there are parentheses $\{ \}$ inside the parentheses $()$: too complicated, we should better keep just one pair of parentheses. Of which kind? Better it should be of the more characteristic kind: the braces $\{ \}$.

For what class of sets $C \subseteq \mathbb{R}$ is the set function μ defined? In other words, for what subsets of the real line is the set $\{X \in C\}$ an *event*? This is a question belonging to the set-theoretic introduction to both measure theory and probability theory; and we have decided not to speak about such questions. Suffice it to say that of course every interval belongs to this class of sets; and that we cannot construct a subset of the real line that does not belong to this class of sets (even if we can *prove* – ineffectively – that such sets C exist).

It turns out that the set function μ_X is necessarily a *measure*.

Indeed, it is, of course, nonnegative (between 0 and 1); as for countable additivity: if $C_1, C_2, \dots, C_n, \dots$ is a sequence of disjoint subsets of the real line (i. e., $C_i \cap C_j = \emptyset$ for $i \neq j$), we have:

$$\{\omega : X(\omega) \in \bigcup_{i=1}^{\infty} C_i\} = \bigcup_{i=1}^{\infty} \{\omega : X(\omega) \in C_i\} \quad (1.5)$$

(using short notations: $\{X \in \bigcup_{i=1}^{\infty} C_i\} = \bigcup_{i=1}^{\infty} \{X \in C_i\}$); these events are disjoint, and from countable additivity of the probability P we obtain:

$$\mu_X\left(\bigcup_{i=1}^{\infty} C_i\right) = P\left(\bigcup_{i=1}^{\infty} \{X \in C_i\}\right) = \sum_{i=1}^{\infty} P\{X \in C_i\} = \sum_{i=1}^{\infty} \mu_X(C_i). \quad (1.6)$$

Two important classes of distributions: discrete distributions given by

$$\mu(C) = \sum_{i: x^i \in C} p(x^i), \quad (1.7)$$

where $x^1, x^2, \dots, x^n, \dots$ are the values taken by the random variable, and the nonnegative numbers $p(x) = P\{X = x\}$ (the sum of *all* $p(x^i)$ is equal to $P\{X \in \{x^1, x^2, \dots, x^n, \dots\}\} = P(\Omega) = 1$); and continuous distributions given by

$$\mu(C) = \int_C p(x) dx, \quad (1.8)$$

where the nonnegative function $p(x) = p_X(x)$ is called the *distribution density*, or the *probability density*, or just *density* (just as $\sum_i p(x^i) = 1$, we have $\int p(x) dx = 1$; if we are not showing the limits within which the variable of summation or of integration changes, we mean that it runs through *all* possible values: in the case of the integral, from $-\infty$ to ∞).

Microtheorem 1.1. *If two random variables X and Y are equivalent: $X \sim Y$, they have the same distribution: $\mu_X = \mu_Y$.*

Proof. We have to prove that for an arbitrary set $C \subseteq \mathbb{R}$ (yes, yes, an arbitrary set belonging to a class that we did not describe precisely, but which is very vast, and contains at least all intervals; but I told you that I won't pay attention to such things)

$$\mu_X(C) = \mu_Y(C), \quad (1.9)$$

or

$$P\{X \in C\} = P\{Y \in C\}. \quad (1.10)$$

Now draw a picture (I cannot draw pictures here) of the sample space Ω and its two subsets (events): $\{X \in C\}$ and $\{Y \in C\}$. We have (look at your picture!):

$$\begin{aligned} |P\{X \in C\} - P\{Y \in C\}| &= |P(\{X \in C\} \setminus \{Y \in C\}) - P(\{Y \in C\} \setminus \{X \in C\})| \\ &\leq P(\{X \in C\} \setminus \{Y \in C\}) + P(\{Y \in C\} \setminus \{X \in C\}) = P(\{X \in C\} \Delta \{Y \in C\}), \end{aligned} \quad (1.11)$$

where Δ is the sign for the symmetric difference of two sets:

$$A \Delta B = (A \setminus B) \cup (B \setminus A) \quad (1.12)$$

(draw another picture, or look at the picture you have drawn).

Clearly,

$$\{X \in C\} \Delta \{Y \in C\} \subseteq \{X \neq Y\} \quad (1.13)$$

(in everyday language rather than in that of sets: one of the events $\{X \in C\}$, $\{Y \in C\}$ occurring, but not the other can only happen if X does not coincide with Y). So we have:

$$|P\{X \in C\} - P\{Y \in C\}| \leq P\{X \neq Y\} = 0, \quad (1.14)$$

so $P\{X \in C\} = P\{Y \in C\}$.

A very important concept is that of the *expectation* $E(X)$ of a random variable X . The way in which this concept is introduced in an elementary probability course is far from satisfactory: the expectation is defined separately for discrete and for continuous random variables; and some things are left without proof. The situation is much better if we build probability theory on the foundation of the theory of measure and integration: it turns out that the expectation is nothing but the *Lebesgue integral* with respect to the probability measure P ; but I decided not to use theory of measure and integration properly speaking, so we are going back to the not very satisfactory situation mentioned above.

Convergence of sequences of random variables. Let $X_1, X_2, \dots, X_n, \dots$ be a sequence of random variables, and Z another random variable (all of them, on the same probability space). What should it mean that $X_n \rightarrow Z$ as $n \rightarrow \infty$?

We have *one* concept of convergence of number sequences; but we have several different types of convergence of random variables. These types are as follows:

- *Convergence in probability.* We say that the sequence X_n converges to Z as $n \rightarrow \infty$ in probability; the notations: $X_n \rightarrow_P Z$ ($n \rightarrow \infty$), or $\lim_{n \rightarrow \infty} (P)X_n = Z$ – if for every positive ε

$$P\{|X_n - Z| < \varepsilon\} \rightarrow 1 \quad (n \rightarrow \infty). \quad (1.15)$$

Of course, it is the same as

$$P\{|X_n - Z| \geq \varepsilon\} \rightarrow 0. \quad (1.16)$$

- *Mean-square convergence.* We say that the sequence X_n converges to Z as $n \rightarrow \infty$ in the mean square (the notations: $X_n \rightarrow^{\text{m.s.}} Z$, or $\text{l.i.m.}_{n \rightarrow \infty} X_n = Z$; l.i.m. is short for *limit in mean*) if

$$\lim_{n \rightarrow \infty} E((X_n - Z)^2) = 0. \quad (1.17)$$

- *Almost-sure convergence.* If we have a sequence of X_n and another random variable Z , we can consider the event

$$\{X_n \rightarrow Z (n \rightarrow \infty)\} = \{\omega: X_n(\omega) \rightarrow Z(\omega) (n \rightarrow \infty)\}, \quad (1.18)$$

and also the opposite event:

$$\{X_n \not\rightarrow Z (n \rightarrow \infty)\}. \quad (1.19)$$

We say that X_n converges to Z almost surely (notation: $X_n \rightarrow^{\text{a.s.}} Z$) if the probability of the event (1.18) is equal to 1 (the probability of the event (1.19) is 0).

The limits in all these senses are determined not uniquely, in general, but *almost* uniquely: say, for convergence in probability: if $X_n \rightarrow_P Z$, and Z' is a random variable that is equivalent to Z , then also $X_n \rightarrow_P Z'$ ($n \rightarrow \infty$). The opposite is also true:

Microtheorem 1.2. *If $X_n \rightarrow_P Z$ and also $X_n \rightarrow_P Z'$, then $Z' = Z$ almost surely (i. e., $Z' \sim Z$).*

Proof. For every positive ε and every natural n we have:

$$\{|Z' - Z| \geq 2\varepsilon\} \subseteq \{|X_n - Z| \geq \varepsilon\} \cup \{|X_n - Z'| \geq \varepsilon\} \quad (1.20)$$

(in the language of everyday wisdom: if two cities Z and Z' are at a distance at least 200 miles from one another, then we, X_n , necessarily are at a distance of at least 100 miles from one of these cities – perhaps from both).

So we have:

$$\begin{aligned} P\{|Z' - Z| \geq 2\varepsilon\} &\leq P(\{|X_n - Z| \geq \varepsilon\} \cup \{|X_n - Z'| \geq \varepsilon\}) \\ &\leq P\{|X_n - Z| \geq \varepsilon\} + P\{|X_n - Z'| \geq \varepsilon\}. \end{aligned} \quad (1.21)$$

The left-hand side does not depend on n , while the right-hand side does; let us take $n \rightarrow \infty$:

$$P\{|Z' - Z| \geq 2\varepsilon\} \leq \lim_{n \rightarrow \infty} [P\{|X_n - Z| \geq \varepsilon\} + P\{|X_n - Z'| \geq \varepsilon\}] = 0. \quad (1.22)$$

Of course the probability in the left-hand side cannot be *less* than 0, so

$$P\{|Z' - Z| \geq 2\varepsilon\} = 0. \quad (1.23)$$

Now taking $\varepsilon \rightarrow 0$, we get

$$P\{Z' \neq Z\} = 0. \quad (1.24)$$

The relations between different kinds of convergence:

- From mean-square convergence convergence in probability follows:

$$X_n \rightarrow^{\text{m.s.}} Z \Rightarrow X_n \rightarrow_P Z \quad (1.25)$$

(the arrow \Rightarrow means that what is standing to the right of it follows from what is to the left; the proof is based on Chebyshev's inequality: $P\{|X_n - Z| \geq \varepsilon\} \leq \frac{E((X_n - Z)^2)}{\varepsilon^2}$). The opposite is not true, generally:

$$X_n \rightarrow_P Z \not\Rightarrow X_n \rightarrow^{\text{m.s.}} Z. \quad (1.26)$$

We should have an example for this. Here it is: the sample space $\Omega = (0, 1]$, the probability P is the length. The random variables X_n are defined as follows:

$$X_n(\omega) = \begin{cases} n & \text{for } 0 < \omega \leq 1/n, \\ 0 & \text{for } 1/n < \omega \leq 1. \end{cases} \quad (1.27)$$

Of course this sequence converges in probability to $Z(\omega) \equiv 0$; but not in the mean square: $E((X_n - Z)^2) = n^2 \cdot \frac{1}{n} + 0 \cdot (1 - 1/n) \rightarrow \infty$ as $n \rightarrow \infty$.

- From almost-sure convergence convergence in probability follows:

$$X_n \rightarrow^{\text{a.s.}} Z \Rightarrow X_n \rightarrow_P Z; \quad (1.28)$$

but convergence in probability does not imply almost-sure convergence:

$$X_n \rightarrow_P Z \not\Rightarrow X_n \rightarrow^{\text{a.s.}} Z. \quad (1.29)$$

However a weaker statement in this direction is true:

$$X_n \rightarrow_P Z \Rightarrow \text{there exists a subsequence } X_{n_k} \text{ such that } X_{n_k} \rightarrow^{\text{a.s.}} (k \rightarrow \infty). \quad (1.30)$$

I don't know whether we'll have an occasion to use this statement.

- Finally, almost-sure convergence does not imply convergence in mean square:

$$X_n \rightarrow^{\text{a.s.}} Z \not\Rightarrow X_n \rightarrow^{\text{m.s.}} Z \quad (1.31)$$

(as the same example (1.27) shows), and mean-square convergence does not imply almost-sure convergence:

$$X_n \rightarrow^{\text{m.s.}} Z \not\Rightarrow X_n \rightarrow^{\text{a.s.}} Z. \quad (1.32)$$

Convergence in probability is the weakest of the three types that I introduced into consideration, and it is sometimes useful as their “common denominator”. For example: a simple microtheorem:

Microtheorem 1.3. *If $X_n \rightarrow^{\text{a.s.}} Z$, and $\text{l.i.m.}_{n \rightarrow \infty} X_n = Z'$, then $Z' = Z$ almost surely.*

Proof. It follows from the almost-sure convergence that

$$X_n \rightarrow_P Z, \quad (1.33)$$

and from the mean-square convergence, that

$$X_n \rightarrow_P Z'; \quad (1.34)$$

and now our statement follows from Microtheorem 1.2.

There were some other things in Lecture 1; but I’ll include them in Lecture note 2.