Lecture note 2. Probability theory. Stochastic processes.

In the same way that we consider random variables taking real values, we can also consider *generalized* random variables, taking values in some other space, Sp, instead of the real line $\mathbb{R}$. For example, we often have to consider *random vectors*, taking values in the space $\mathrm{Sp} = \mathbb{R}^r$. A generalized random variable is defined as a function $X: \Omega \mapsto \mathrm{Sp}$ such that, for some specified class of subsets $C \subseteq \mathrm{Sp}$ we can consider the event $\{X \in C\}$: in other words,

$$\{\omega\colon X(\omega) \in C\} \in \mathcal{F}. \tag{2.1}$$

How should this class of sets C be chosen? In the case of real-valued random variables we took all *intervals*; in the case of random vectors, that is, $\mathrm{Sp} = \mathbb{R}^r$, we can take either the class of all r-dimensional "intervals", i.e., r-dimensional parallelepipeds:

$$C = I_1 \times I_2 \times \ldots \times I_r, \tag{2.2}$$

where I_i are intervals; or, with the same result (leading to an equivalent definition), we can take the class of all open r-dimensional sets C.

The distribution of a generalized random variable is defined the same way as for real-valued ones: it is a measure on the space Sp defined by

$$\mu(C) = \mu_X(C) = P\{X \in C\} \tag{2.3}$$

(and the class of all C's that we can consider here is so vast that we cannot produce an example of a subset $C \subseteq \mathrm{Sp}$ that cannot be put into the function (2.3) – even if such sets C exist).

If $X_1, \ldots, X_r$ are random variables (of course, on the same probability space $(\Omega, \mathcal{F}, P)$), their *joint distribution* $\mu = \mu_{X_1, \ldots, X_r}$ is, by definition, the distribution of the r-dimensional random vector $\boldsymbol{X}$ with components $X_1, \ldots, X_r$: for an r-dimensional set C

$$\mu(C) = \mu_{X_1, \ldots, X_r}(C) = P\{(X_1, \ldots, X_n) \in C\}. \tag{2.4}$$

Just as in the case of one-dimensional distributions, we can consider r-dimensional discrete distributions given by

$$\mu(C) = \sum_{(x_1, \ldots, x_r) \in C} p(x_1, \ldots, x_n) \tag{2.5}$$

(compare formula (1.17)), and continuous distributions with an r-dimensional density (the joint probability density of the random variables $X_1, \ldots, X_r$):

$$\mu(C) = \int \cdots \int_C p(x_1, \ldots, x_n)\, dx_1 \ldots dx_n. \tag{2.6}$$

You are supposed to know the multidimensional Gaussian (normal) distribution.

A very important question about random variables is whether limit passage under the expectation sign is possible; that is: given that $X_n \to Z$ in some sense, does it follow that $E(X_n) \to E(Z)$?

For mean-square convergence, the answer is **yes**, provided that the expectations in question exist:

$$|E(X_n) - E(Z)| = |E(X_n - Z)| \leq \sqrt{E\big((X_n - Z)^2\big)} \to 0 \qquad (n \to \infty). \tag{2.7}$$

The inequality holds because

$$0 \leq \mathrm{Var}(X_n - Z) = E\big((X_n - Z)^2\big) - \big(E(X_n - Z)\big)^2, \quad \big(E(X_n - Z)\big)^2 \leq E\big((X_n - Z)^2\big). \tag{2.8}$$

But

$$X_n \to_P Z \;\not\Rightarrow\; E(X_n) \to E(Z), \tag{2.9}$$

$$X_n \to^{\text{a. s.}} Z \;\not\Rightarrow\; E(X_n) \to E(Z), \tag{2.10}$$

even under the assumption that the expectations $E(X_n)$, $E(Z)$ exist. This can be seen, once more, from example (1.27).

So the question should be not "given that $X_n \to Z$ almost surely or in probability, does it follow that $E(X_n) \to E(Z)$," but rather "given that $X_n \to Z$ almost surely or in probability, *under what supplementary conditions* can we guarantee that $E(X_n) \to E(Z)$?"

In this direction I'll formulate three big theorems. (They are, in fact, theorems in the theory of measure and integration, and I'll give them without proofs – formulated with no mention of measure or Lebesgue integral.)

Theorem 2.1 (about monotone convergence). *Let* $0 \leq X_1 \leq X_2 \leq \ldots \leq X_n \leq \ldots$ *be a non-decreasing sequence of nonnegative random variables* (the inequalities are supposed to be satisfied for all $\omega \in \Omega$). Of course, there exists the limit

$$Z = Z(\omega) = \lim_{n \to \infty} X_n(\omega), \tag{2.11}$$

possibly infinite for some ω. The limit, Z, is a (generalized) random variable, taking values in the extended right half-line $[0, \infty]$.

Then

$$E(X_n) \to E(Z) \qquad (n \to \infty). \tag{2.12}$$

Here we should make it clear what we mean by the expectation $E(Z)$ of the generalized random variable Z – taking, possibly, at some sample points the value ∞.

If $P\{Z = \infty\} > 0$, we take, by definition, $E(Z) = \infty$.

And if $P\{Z = \infty\} = 0$ (i. e., if Z is finite almost surely), we handle Z the same way as we do usual number-valued random variables.

Theorem 2.2 (about dominated convergence). *If all* X_n, *for all* ω, *are dominated in absolute value by some positive random variable* Y *having a finite expectation*:

$$|X_n| \leq Y \text{ for all } n, \qquad E(Y) < \infty, \tag{2.13}$$

and $X_n \to Z$ *almost surely, or in probability, then*

$$E(X_n) \to E(Z) \qquad (n \to \infty). \tag{2.14}$$

In the books on measure and integration theory this theorem is usually formulated with almost-sure (or, in the language of measure theory, *almost-everywhere*) convergence; but it is true also for convergence in probability (in *measure*).

Theorem 2.3 (Fatou's Lemma). *Let* $X_1, X_2, \ldots, X_n, \ldots$ *be a sequence of nonnegative random variables.* We don't suppose that the limit $\lim_{n\to\infty} X_n(\omega)$ exists for all – or *almost* all – or even for any ω; but always exist the upper limit $\overline{\lim}_{n\to\infty} X_n(\omega)$ and the lower limit $\underline{\lim}_{n\to\infty} X_n(\omega)$, that may be equal to ∞ for some ω's.

We have:

$$E(\underline{\lim}_{\,n\to\infty} X_n) \le \underline{\lim}_{\,n\to\infty} E(X_n). \tag{2.15}$$

(In the lecture, I showed how one can remember the direction of the sign $\le$ and whether it is the upper or the lower limit – using example (1.27) and another example.)

Theorems 2.1 and 2.3 are definitely about number-valued random variables: otherwise we couldn't speak of inequalities between them; but Theorem 2.2 remains true for sequences of random vectors – if we take instead of the absolute values $|X_n|$ the lengths of the random vectors.

Now we go to random functions and stochastic processes.

A *random function* is a family of random variables $X_t = X_t(\omega)$ taking values in some space Sp and depending on a parameter t running over some set T.

If the parameter set T is a part of the real line: $T \subseteq \mathbb{R}$; *if* we interpret the parameter t as time; and *if* we interpret X_t as motion of some random point in the space Sp, we call the random function X_t a *stochastic process*.

Example 2.1 (not given in Lecture 2). Let X_1, Y_1, X_2, Y_2 be independent random variables having each the normal distribution with parameters $(0, 1)$. For every real t, let

$$Z_t = X_1 \cos t + Y_1 \sin t + X_2 \cos 2t + Y_2 \sin 2t. \tag{2.16}$$

The stochastic process Z_t is determined by finitely many (namely, four) random variables. The stochastic processes that we are going to face later are not of this kind: they cannot be determined by finitely many random variables, and even if they can, their representation through these random variables may be so complicated that it is of no use.

The types of convergence that were introduced for sequences of random variables are introduced in the same way for random functions: for example, we say that $X_t \to_P Z$ as $t \to t_0$ (or: $\lim_{t\to t_0}(P) X_t = Z$) if for every positive ε we have $P\{|X_t - Z| < \varepsilon\} \to 1$ as $t \to t_0$ (or: $\lim_{t\to t_0} P\{|X_t - Z| \ge \varepsilon\} = 0$). The properties of different types of limit remain the same in the case of continuous parameter.

A random function is a function $X_t(\omega)$ of two variables: t and ω.

If we fix the first variable, it is a function of the argument ω only: a random variable. So we can consider for X_t's things that are usually considered for random variables: in particular, their distributions; and their joint distributions.

For arbitrary n points t_1, ..., $t_n \in T$, we can consider the joint distribution $\mu_{X_{t_1}, ..., X_{t_n}}$ of the random variables X_{t_1}, ..., X_{t_n}:

$$\mu_{X_{t_1}, ..., X_{t_n}}(C) = P\{(X_{t_1}, ..., X_{t_n}) \in C\}, \qquad C \subseteq \mathrm{S\small P}^n. \tag{2.17}$$

These distributions are called the *finite-dimensional distributions* corresponding to the random function X_t, $t \in T$. We are going to use for them a shorter notation:

$$\mu_{X_{t_1}, ..., X_{t_n}} = \mu_{t_1, ..., t_n}. \tag{2.18}$$

On the other hand, we can consider the random function $X_t(\omega)$ *for fixed sample point* ω, and then it is a function of the argument t only. This function is called a *sample function*; or a *realization* of the random function X_t; in the case that we call X_t a *stochastic process*, for a sample function the term *trajectory* is also used.

E. g., all trajectories of the stochastic process of Example 2.1 are periodic functions with period 2π (but not all periodic functions are trajectories of this process).

As for an example of the system of finite-dimensional distributions, I did not give any in Lecture 2; but in some future lecture note I'll give some problems that will provide such examples.

Now some simple piece of theory about the system of finite-dimensional distributions.

Microtheorem 2.1. *Let* $\mu_{t_1, ..., t_n}$, t_1, ..., $t_n \in T$, $n = 1, 2, 3, ...$, *be the system of finite-dimensional distributions of some random function* X_t, $t \in T$, *taking values in a space* Sp.

Then the following two conditions (called *consistency conditions*) *are satisfied*:

1) *for every permutation* $(i_1, i_2, ..., i_n)$ *of the numbers* $(1, 2, ..., n)$, *every* n*-tuple of elements* $t_i \in T$ *and of sets* $C_i \subseteq$ Sp *we have*:

$$\mu_{t_{i_1}, t_{i_2}, ..., t_{i_n}}(C_{i_1} \times C_{i_2} \times C_{i_n}) = \mu_{t_1, t_2, ..., t_n}(C_1 \times C_2 \times ... \times C_n); \tag{2.19}$$

2) *for all* t_1, t_2, ..., $t_n \in T$ *and all* C_1, C_2, ..., $C_{n-1} \subseteq$ Sp *we have*:

$$\mu_{t_1, ..., t_{n-1}, t_n}(C_1 \times C_2 \times ... \times C_{n-1} \times \mathrm{S\small P}) = \mu_{t_1, ..., t_{n-1}}(C_1 \times C_2 \times ... \times C_{n-1}). \tag{2.20}$$

The consistency conditions are formulated in terms of the values of distribution measures on multidimensional "parallelepipeds".

The **proof** is very simple: both sides in formula (2.19) are, by definition, the probability of the event

$$\{X_{t_1} \in C_1, X_{t_2} \in C_2, ..., X_{t_n} \in C_n\}, \tag{2.21}$$

but in the left-hand side the order of mentioning X_{t_k} is different; and both sides in (2.20) are the probability

$$P\{X_{t_1} \in C_1, X_{t_2} \in C_2, ..., X_{t_{n-1}} \in C_{n-1}\}, \tag{2.22}$$

but in the left-hand side the description of this event includes the condition $X_{t_n} \in$ Sp is added, which, of course, is no restriction, because all $X_{t_n}(\omega)$'s belong to the space Sp.

Let us see how the consistency conditions (2.19), (2.20) look for the two important classes of distributions: discrete and continuous.

For discrete distributions, let us consider multidimensional "probability mass functions" $p_{t_1, t_2, ..., t_n}(x_1, x_2, ..., x_n)$ (being equal to the probability $P\{X_{t_1} = x_1, X_{t_2} = x_2, ..., X_{t_n} = x_n\}$), so that for $C \subseteq \mathrm{Sp}^n$

$$\mu_{t_1, t_2, ..., t_n}(C) = \sum_{(x_1, x_2, ..., x_n) \in C} p_{t_1, t_2, ..., t_n}(x_1, x_2, ..., x_n). \tag{2.23}$$

Conditions 1), 2) are rewritten in the following form:

$$p_{t_{i_1}, t_{i_2}, ..., t_{i_n}}(x_{i_1}, x_{i_2}, ..., x_{i_n}) = p_{t_1, t_2, ..., t_n}(x_1, x_2, ..., x_n), \tag{2.24}$$

$$\sum_{x_n \in \mathrm{Sp}} p_{t_1, ..., t_{n-1}, t_n}(x_1, ..., x_{n-1}, x_n) = p_{t_1, ..., t_{n-1}}(x_1, ..., x_{n-1}). \tag{2.25}$$

For continuous distributions with densities $p_{t_1, t_2, ..., t_n}(x_1, x_2, ..., x_n)$:

$$\mu_{t_1, t_2, ..., t_n}(C) = \int \cdots \int_C p_{t_1, t_2, ..., t_n}(x_1, x_2, ..., x_n)\, dx_1\, dx_2 \ldots dx_n, \tag{2.26}$$

the consistency conditions are rewritten in the form

$$p_{t_{i_1}, t_{i_2}, ..., t_{i_n}}(x_{i_1}, x_{i_2}, ..., x_{i_n}) = p_{t_1, t_2, ..., t_n}(x_1, x_2, ..., x_n), \tag{2.27}$$

$$\int_{\mathrm{Sp}} p_{t_1, ..., t_{n-1}, t_n}(x_1, ..., x_{n-1}, x_n)\, dx_n = p_{t_1, ..., t_{n-1}}(x_1, ..., x_{n-1}). \tag{2.28}$$

Of course, since a probability density is not defined in a unique way, but only *almost* uniquely (a density has many *versions*, but all of them differ only on sets that can be disregarded by integration), the precise formulation should be: equalities (2.27), (2.28) are satisfied except for a set of points $(x_1, x_2, ..., x_n)$ (or $(x_1, ..., x_{n-1})$) that can be disregarded by integration; or: one *version* of the densities is such that (2.27), (2.28) are satisfied.