Introduction to probability theory in the Discrete Mathematics course

Jiří Matoušek (KAM MFF UK)

Version: Oct/18/2013

Introduction

This detailed syllabus contains definitions, statements of the main results and concise comments. It should not substitute a textbook, and you will find no proofs and no solved exercises here. A part of the material is covered in more detail in the book J. Matoušek, J. Nešetřil: *Invitation to Discrete Mathematics*, 2nd edition, Oxford University Press, 2009 (Chap. 10). There are many textbooks of probability theory; an introduction for beginners is, e.g., H. Tijms: *Understanding Probability.* Cambridge Univ. Press., 2004, and a more advanced and comprehensive book is G. R. Grimmett, D. R. Stirzaker: *Probability and Random Processes*, Oxford University Press, 2001. A reasonable-looking textbook at `http://www.math.uiuc.edu/~r-ash/BPT/BPT.pdf` can even be freely downloaded at present.

1. Chance, or randomness, is often symbolized by a die (plural: dice). The probability of getting a six in a single roll is $\frac{1}{6}$. The probability of two sixes in two rolls is $\frac{1}{36}$. Tossing a fair coin is still a simpler experiment: both heads and tails have probability $\frac{1}{2}$.

2. What is meant by probability? For a "real" probability this is a philosophical question, with various answers, none of them completely satisfactory. (Dice can be rolled repeatedly, "empirical probability" can be defined as a limit for an enormous number of rolls—but does it make any sense to talk about the probability of unique events, such as the gorillas, or the humans, surviving the year 2015?) Where does randomness come from? (Maybe from quantum theory? Or from chaos theory—what we in our ignorance perceive as random might perhaps be really predetermined?)

3. The mathematical theory of probability does not consider these philosophical conundrums. It construct a mathematical *model* of probability. This model works amazingly well, ask insurance companies (one that are not going bankrupt, that is). But for real-life problems it must be applied thoughtfully (ask insurance companies that are going bankrupt). It is well known to psychologists that people generally cannot estimate probabilities reliably, especially very small or very large ones (see, for example, *Thinking fast and slow* by Daniel Kahneman). For example, in an experiment, only 45 % students were able to finish their thesis by the time for which they were 99 % sure to finish it.

4. Probability in computer science: randomized algorithms (simpler and faster than deterministic ones; often the only feasible approach for massive amounts of data), statistical tests and evaluation of experiments, mathematical proofs.

Probability spaces, mainly discrete ones

5. We perform a random experiment. Let Ω denote the set of all possible outcomes. Examples:

   - Tossing a coin: $\Omega = \{H, T\}$ (heads, tails).
   - Rolling a die: $\Omega = \{1, 2, 3, 4, 5, 6\}$.
   - Three consecutive coin tosses:
     $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$.
   - The first raindrop falling on a square garden table: $\Omega =$ all points of a square.

   Elements of Ω are called **elementary events**.

6. An **event** is a subset $A \subseteq \Omega$. Examples:

   - coin toss—we can consider 4 events: $A_1 = \{H\}$ (heads tossed), $A_2 = \{T\}$ (tails tossed), $A_3 = \{H, T\}$ (heads *or* tails tossed, this is a **certain event**, it always occurs), and $A_4 = \emptyset$ (nothing tossed, this is the **impossible event**, it never occurs).
   - rolling a die, $\Omega = \{1, 2, 3, 4, 5, 6\}$: "an even number rolled" $A_1 = \{2, 4, 6\}$, "six rolled" $A_2 = \{6\}$, "nothing rolled" $A_3 = \emptyset$.
   - first raindrop: "falls in the left half of the table", "falls at most 10 cm from the border", "falls on a flower printed on the tablecloth", etc.

7. To every event A we assign a real number $\mathrm{P}[A] \in [0, 1]$, called the **probability of A**. Examples:

   - For tossing a fair coin we have
     $\mathrm{P}[\{H\}] = \mathrm{P}[\{T\}] = \frac{1}{2}$, $\mathrm{P}[\{H, T\}] = 1$, $\mathrm{P}[\emptyset] = 0$.
   - For a "fair" die we have $\mathrm{P}[\{1\}] = \mathrm{P}[\{2\}] = \cdots = \mathrm{P}[\{6\}] = \frac{1}{6}$, and, for example, $\mathrm{P}[\{1, 3, 5\}] = \frac{1}{2}$.
   - We can also consider a loaded die (of a cheater) with $\mathrm{P}[\{6\}] = \frac{1}{5}$, $\mathrm{P}[\{1\}] = \mathrm{P}[\{2\}] = \cdots = \mathrm{P}[\{5\}] = \frac{4}{25}$. Thus, not all one-element events need to have the same probability.

8. A *probability space* (also sometimes called a *sample space*) is formally a triple $(\Omega, \mathcal{F}, \mathrm{P})$, where

   - Ω tells us, which elementary events are considered;
   - $\mathcal{F} \subseteq 2^\Omega$ specifies, which subsets of Ω are admitted as events;
   - and $\mathrm{P}\colon \mathcal{F} \to [0, 1]$ is a function that assigns to every event its probability.

9. Every probability space has to satisfy certain axioms. Here we will not present them in full generality, since we do not yet have the necessary mathematical tools ready (specifically, contemporaty probability is based in *measure theory*, which is usually covered only later in the curriculum, if at all). We introduce only *discrete* probability spaces precisely, ones with the set Ω finite or countable, and such that all subsets of Ω are events.

We are thus leaving aside, e.g., "geometric" probability (what is the probability of the first raindrop falling at most 10 cm from the border, etc.), as well as probabilities concerning arbitrarily long sequences. (For example: We repeatedly toss a coin. Is it more likely that we first encounter the sequence HHT, or HTT? This is a "real" question, you can make bets about it! And it has a surprising answer: HHT wins twice as often.)

10. A **discrete probability space** is a triple $(\Omega, \mathcal{F}, \mathrm{P})$, where Ω is a finite or countable set, $\mathcal{F} = 2^\Omega$ (that is, every subset is an event), the probability of every event $A \subseteq \Omega$ satisfies

$$\mathrm{P}[A] = \sum_{\omega \in A} \mathrm{P}[\{\omega\}],$$

and $\mathrm{P}[\Omega] = 1$.

This means that a discrete probability space is fully determined by the probabilities of all one-element events. The probabilities of these singletons can be chosen as arbitrary nonnegative numbers whose sum over Ω equals 1 (for infinite Ω this is a sum of an infinite series).

11. Here we will work mainly with *finite* discrete probability spaces, where Ω is a finite set. (In this case we will usually omit the word "discrete".)

12. A basic example of a finite probability space is a *classical probability space*, where $\mathrm{P}[A] = \frac{|A|}{|\Omega|}$; thus, all singleton events have the same probability.

13. Remark: If we wanted every singleton to have the same probability for Ω infinite, these probabilities would have to be all 0. Indeed, in geometric probability (a random point in a square, say), every single point has zero probability. At the same time, there are events, such as the whole square, with nonzero probability, which looks paradoxical—how can infinitely many points with zero probability combine into something with nonzero probability? This is where measure theory is needed, to define geometric probability in a consistent way.

14. Here are some specific examples of classical probability spaces, particularly important for discrete mathematics.

    - A random sequence of n tosses by a fair coin: $\Omega = \{H, T\}^n$, every possible sequence of outcomes has probability $1/2^n$.
    - A random permutation of the set $\{1, 2, \ldots, n\}$: $\Omega = S_n$ (the set of all permutations), each permutation has probability $1/n!$.

15. Let us stress that for a given mathematical or practical problem, a probability space can often be chosen in several different ways; some of them may turn out to be more convenient than others.

16. What are probability spaces good for? They provide a safe foundation and often they help in clarifying various tricky questions. Example: *Bertrand's box paradox*; here we present it in a version with cards. There are three cards in a hat: one with both sides red, another with both sides black, and a third one with one side red and one black. You

blindly draw one of the cards and put it on the table. Then you look and see that the top side is red. What is the probability that the bottom side is red too? An intuitive answer is $\frac{1}{2}$, but the truth is $\frac{2}{3}$. (Also see a modern version— the *Monty Hall problem*).

Conditional probability, independent events

17. The **conditional probability** of an event A assuming that an event B has occurred is defined as
$$\mathrm{P}[A \mid B] = \frac{\mathrm{P}[A \cap B]}{\mathrm{P}[B]}$$
(it is defined only for $\mathrm{P}[B] > 0$).

18. Often it is useful to compute the probability of A by "case distinction": if $B_1, \ldots, B_n$ are *disjoint* events whose union is all of Ω, then
$$\mathrm{P}[A] = \mathrm{P}[A \mid B_1]\,\mathrm{P}[B_1] + \mathrm{P}[A \mid B_2]\,\mathrm{P}[B_2] + \cdots + \mathrm{P}[A \mid B_n]\,\mathrm{P}[B_n]\,.$$
This simple statement is sometimes called the **theorem about complete probability**.

19. Example (somewhat artificial but valuable). Let us say that 0.1% of the population is HIV-infected. Let us imagine that you had a HIV test, which is known to give a positive result for an *actually infected* person in 95 % of cases, while for persons *not infected* it gives a (false) positive result in 5 %. You have no particular reason to expect to be infected (i.e., you can be regarded as a "random person" in this respect), but your test comes out positively. What is the probability, that you are really infected? Here the probability space is the population (and you are a randomly selected person), the event $H =$ HIV positive, the event $T =$ tests positively, assumptions: $\mathrm{P}[H] = 0.001$, $\mathrm{P}[T \mid H] = 0.95$, $\mathrm{P}[T \mid \text{not } H] = 0.05$; we want to know: $\mathrm{P}[H \mid T]$. The result is less than 2 %, so you need not worry too much.

By formulating the reasoning in this example in a general way, we obtain the **Bayes theorem**:
$$\mathrm{P}[B_i \mid A] = \frac{\mathrm{P}[A \mid B_i]\,\mathrm{P}[B_i]}{\sum_{j=1}^{n} \mathrm{P}[A \mid B_j]\,\mathrm{P}[B_j]},$$
where $B_1, \ldots, B_n$ are events as in the previous item (disjoint and covering Ω).

20. Two events A and B are called **independent** if $\mathrm{P}[A \cap B] = \mathrm{P}[A]\,\mathrm{P}[B]$. Equivalently, A and B are independent if $\mathrm{P}[B] = 0$ or $\mathrm{P}[A \mid B] = \mathrm{P}[A]$. (Intuition: if we learn whether event B has occurred, we have gained no new information about the probability of B occurring.) More generally: events $A_1, A_2, \ldots, A_n$ are independent if we have, for every index set $I \subseteq \{1, 2, \ldots, n\}$, the equality $\mathrm{P}[\bigcap_{i \in I} A_i] = \prod_{i \in I} \mathrm{P}[A_i]$.

21. Example: we consider a random sequence of n zeros and ones (or n fair coin tosses),
$A_1 = \{$the first 5 entries are all 1s$\}$,
$A_2 = \{$the sixth entry is 0$\}$,
$A_3 = \{$there are an odd number of 1s among the last 5 entries$\}$.
The events A_1 and A_2 are "obviously" independent. The events A_1, A_2, A_3 are also independent, but this has to be checked carefully according to the definition.

Another example: a random permutation π of $\{1, 2, \ldots, 32$ (a well-mixed pack of cards); the events $A = \{\pi(1) = 1\}$ and $B = \{\pi(2) = 2\}$ are *not* independent!

22. There is a widespread error in intuitive interpretations of independent events ("I have not rolled a six for such a long time—so now it *must* come!"—in reality, the dice have no memory and the probability does not depend on what happened before).

23. Example: $\Omega = \{$blue-eyed blond woman, brown-eyed brown-haired woman, blue-eyed brown-haired man, brown-eyed blond man$\}$, all four have probability $\frac{1}{4}$ (classical probability space); events $A_1 =$ blue eyes, $A_2 =$ blond hair, $A_3 =$ woman (or, if you prefer numbers, $\Omega = \{000, 011, 110, 101\}$). Every two of these events are independent, but all three are not.

24. **Product of (discrete) probability spaces:**
$(\Omega_1, 2^{\Omega_1}, \mathrm{P}_1) \times (\Omega_2, 2^{\Omega_2}, \mathrm{P}_2) = (\Omega, 2^{\Omega}, \mathrm{P})$,
where $\Omega = \Omega_1 \times \Omega_2$ and $\mathrm{P}[A] = \sum_{(\omega_1,\omega_2)\in A} \mathrm{P}_1[\{\omega_1\}]\mathrm{P}_2[\{\omega_2\}]$.

25. Let us consider a product probability space with $\Omega = \Omega_1 \times \Omega_2$, and let A_1 be an event that depends only on the first component; that is, $A_1 = A_1' \times \Omega_2$ for some $A_1' \subseteq \Omega_1$. Similarly, let A_2 be an event that depends only on the second component. Then A_1 and A_2 are independent.

26. Example: a "sequence of n tosses of a fair coin" can be constructed as a product probability space, the product of n spaces, one for each coin toss. A more interesting example: $\Omega = \{0, 1\}$, $\mathrm{P}[\{1\}] = p$ (an experiment with success probability p), the product of n copies of this space models a sequence of n independent repetitions of this experiment. For $\omega \in \{0, 1\}^n$ we have $\mathrm{P}[\{\omega\}] = p^j(1-p)^{n-j}$, where j is the number of 1s in ω.

Random variables, expectation

27. A **(real) random variable** on a discrete probability space $(\Omega, 2^{\Omega}, \mathrm{P})$ is an arbitrary function $X\colon \Omega \to \mathbf{R}$.

Example: We toss a fair coin n times, so $\Omega = \{H, T\}^n$, i.e., all n-term sequences of H's and T's. Examples of random variables on this space: the number of times heads was tossed; the number of heads minus the number of tails, the sine of the number of heads. (Remark: One can also consider complex random variables, or random variables whose values are points in the plane, etc. But here we talk only about real random variables.) Be very careful to distinguish between an *event* (only two possible outcomes, occurred/not occurred—formally an event is a set), and a *random variable* (the result is a number, and formally a random variable is a function). In a sense, an event can be regarded as a special case of a random variable, as will be mentioned later.

28. Let $X\colon \Omega \to \mathbf{R}$ be a random variable on a discrete probability space $(\Omega, 2^{\Omega}, \mathrm{P})$. The **expectation of X** is defined as

$$\mathbf{E}[X] := \sum_{\omega\in\Omega} \mathrm{P}[\{\omega\}]\, X(\omega).$$

(For Ω infinite, the infinite series in the definition of $\mathbf{E}[X]$ need not converge, and then the expectation of X does not exist.) Intuition: $\mathbf{E}[X]$ is the average value of X if we make a very large number of experiments.

29. $\mathbf{E}[X]$ can also be computed as $\sum_{v \in V} v \cdot \mathrm{P}[X = v]$, where V is the set of all values attained by X (this is just rearranging the sum in the definition). Here $\mathrm{P}[X = v]$ is a commonly used shorthand for the probability of the event $\{\omega \in \Omega : X(\omega) = v\}$.

30. Another notion, which in practice is sometimes confused with the expectation: the **median** of a real random variable X is a number m such that $\mathrm{P}[X < m] \leq \frac{1}{2}$ and $\mathrm{P}[X > m] \leq \frac{1}{2}$. For example, half of people have salary smaller than the median and half larger. But, as a rule, a substantial majority has salary smaller than average (and everyone wants at least the average salary—which many politicians are willing to promise before elections).

31. *Linearity of expectation*: $\mathbf{E}[\alpha X] = \alpha \mathbf{E}[X]$, $\mathbf{E}[X + Y] = \mathbf{E}[X] + \mathbf{E}[Y]$ for any two, totally arbitrary, random variables X and Y, and $\alpha \in \mathbf{R}$ (assuming that all the expectations exist). The proof is immediate from the definition. But, for example, $\mathbf{E}[XY] \neq \mathbf{E}[X]\,\mathbf{E}[Y]$ in general.

32. The **indicator** of an event A is the random variable $I_A \colon \Omega \to \{0, 1\}$ given by

$$I_A(\omega) = \begin{cases} 1 & \text{for } \omega \in A \\ 0 & \text{for } \omega \notin A. \end{cases}$$

    We have $\mathbf{E}[I_A] = \mathrm{P}[A]$. The indicator I_A is the counterpart of A in the realm of random variables; in this sense, random variables are more general than events.

33. Examples (calculating the expectation):

    - $X =$ number of heads in a sequence of n tosses of a fair coin. The expectation $\mathbf{E}[X]$ can be calculated according to the definition. A much easier way, the *method of indicators*: A_i is the event "the ith comes out heads"; $X = I_{A_1} + I_{A_2} + \cdots + I_{A_n}$, $\mathbf{E}[I_{A_i}] = \frac{1}{2}$, $\mathbf{E}[X] = \frac{n}{2}$.
    - $Y =$ the number of left maxima of a random permutation π of the set $\{1, 2, \ldots, n\}$, i.e., the number of i such that $\pi(j) < \pi(i)$ for all $j < i$. Event $A_i :=$ "i is a left maximum". We calculate $\mathbf{E}[Y] = 1 + \frac{1}{2} + \cdots + \frac{1}{n} \approx \ln n$.

34. Example: Every graph $G = (V, E)$ has a bipartite subgraph with at least $\frac{1}{2}|E|$ edges. Proof: Partition the vertex set into two parts at random, let X be the number of edges going "across", and calculate $\mathbf{E}[X] = \frac{1}{2}|E|$.

35. Random variables X, Y (on the same probability space) are called **independent** if for every $a, b \in \mathbf{R}$, the events "$X \leq a$" and "$Y \leq b$" are independent. Similarly for more than 2 random variables.

36. For *independent* random variables we have $\mathbf{E}[XY] = \mathbf{E}[X]\,\mathbf{E}[Y]$.