CHAPTER 0

Basics of probability theory

1. Introduction

In the next few weeks, we will study the so-called "probabilistic method". This is the – sometimes surprising – application of probability theory to problems in discrete mathematics, or the design of algorithms.

Probability theory is a very wide area, but in this course we will only need a smaal subset of it. This chapter serves as a brief recapitulation of those elements of (discrete) probability theory that we will encounter in the course.

For a more thorough introduction to probability theory, see e.g. the books by Feller [Fel68a; Fel68b], or Billingsley [Bil95].

2. Probability

2.1. Deterministic vs. random. Many phenomena in nature can be described deterministically. For example, Hooke's law tells us that a spring, that is on one side attached to a fixed object, and on the other side is pulled by a certain force, will be displaced by a distance that is proportional to that force. Whenever we pull the same string by the same force, we will find that is displaces by the same amount.

Other experiments have an uncertain outcome. For example, one cannot easily predict the outcome of a coin toss or die roll, based on previous experiments. Although in principle one should be able to compute the outcome of a die roll when all the influencing factors are precisely known, for all practical purposes, the outcome is random.

2.2. Probability spaces. Probability theory is often introduced axiomatically. The starting point is a (potentially uncountable) set Ω, called the **sample space**. The sample space Ω is equipped with a so-called sigma-algebra $\mathcal{F}$, on which a probability measure is defined. A sigma-algebra $\mathcal{F}$ is a collection of subsets of the sample space satisfying the following properties:

(i) $\mathcal{F}$ is non-empty;
(ii) $\mathcal{F}$ is closed under taking complements, i.e. if $A \in \mathcal{F}$, then $\bar{A} \equiv \Omega \setminus A \in \mathcal{F}$;
(iii) $\mathcal{F}$ is closed under countable unions, i.e. if $A_1, A_2, \ldots \in \mathcal{F}$, then $\bigcup_i A_i \in \mathcal{F}$.

In this course, our sample space will always be finite or countably infinity, and we will always take he sigma-algebra to be the full power set of Ω. Hence, we will always suppress reference to the underlying sigma-algebra; we include its definition here just for the sake of completeness.

Elements of the sigma-algebra are called **events** (so, in our situation, an event is just any subset of the sample space Ω).

The sample space is just the set of all possible outcomes of a certain random experiment. For example, a sample space associated with a die roll is $\Omega = \{1, 2, 3, 4, 5, 6\}$, while an obvious choice for the sample space of a single coin toss is $\Omega = \{H, T\}$. Events have a nice interpretation as well; when rolling a die, the event $\{2, 4, 6\}$ means: rolling an even number.

Often, sample spaces and events will be much more elaborate. For example, a sample space for rolling a die repeatedly until heads comes up, is given by $\Omega = \{H, TH, TTH, TTTH, \ldots\}$. In this case, the sample space is countably infinite. The event $\{TH, TTTH, TTTTTH, \ldots\}$ is interpreted as: an odd number of tails is rolled before heads is rolled.

Set theory is a useful language for events, as the set-theoretic operators $\cup$ (union), $\cap$ (intersection), resp. $\bar{\cdot}$ (complement) correspond to the logical operators OR, AND, resp. NOT.

A function $\mathbb{P} : \mathcal{F} \to [0,1]$ (sending events to real numbers in the interval $[0,1]$) is called a probability measure if it satisfies the following conditions:

(i) $\mathbb{P}(\Omega) = 1$;
(ii) If $\{A_1, A_2, \ldots\}$ is a countable collection of pairwise disjoint events, then $\mathbb{P}(\bigcup A_i) = \sum \mathbb{P}(A_i)$.

For any event A, the quantity $\mathbb{P}(A)$ is to be interpreted as a probability. Then the first condition simply states that the probability of the sure event Ω is 1, while the second condition states that the probability of a union of pairwise disjoint events is simply the sum of their probabilities.

There are a number of properties of probability measures that follow from the definition.

(i) $\mathbb{P}(\emptyset) = 0$;
(ii) For all events A, $\mathbb{P}(\bar{A}) = 1 - \mathbb{P}(A)$.
(iii) **Union bound**: For a sequence of events $\{A_1, A_2, \ldots\}$, $\mathbb{P}(\bigcup_i A_i) \leq \sum_i \mathbb{P}(A_i)$.

EXAMPLE. *Consider a finite state space $\Omega = \{1, 2, \ldots, n\}$. Let $p_1, p_2, \ldots, p_n$ be real numbers satisfying $p_i \geq 0$ for all $i = 1, 2, \ldots, n$ and $\sum_{i=1}^{n} p_i = 1$, then $\mathbb{P}(\{i\}) = p_i$ defines a probability measure.*

2.3. Random variables. A **random variable** is a function[1] $\mathbf{X} : \Omega \to \mathbb{R}$. Note that random variables form a real vector space.

EXAMPLE. *Suppose that we are betting €1 that a coin toss comes up heads. We let $\mathbf{X}$ be the pay-off of this bet, i.e. $\mathbf{X} = 1$ if the coin comes up heads, and $\mathbf{X} = -1$ if the coin comes up tails.*

We shall often write $\{\mathbf{X} \leq x\}$ (and similar expressions) as a shorthand for the event $\{\omega \in \Omega : \mathbf{X}(\omega) \leq x\}$, and we write $\mathbb{P}(\mathbf{X} \leq x)$ for the probability of this event.

When $\mathbf{X}$ takes values in a countable set, it is called a **discrete** random variable. In this course, we will work exclusively with discrete random variables. If $\mathbf{X}$ is a discrete random variable, then its density function is

$$p(x) = p_X(x) = \mathbb{P}(\mathbf{X} = x).$$

The **indicator random variable** $\mathbb{1}_A$ of an event A is the random variable

$$\mathbb{1}_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases}.$$

EXAMPLE. *As a more elaborate example, that fits more to the contents of the course, suppose that we start out with the complete graph on n vertices. For each of the $\binom{n}{2}$ edges in this graph, we toss a coin, and remove the edge if it comes up heads. In this setting, both the number of edges in the resulting graph, and the number of isolated vertices in the resulting graph are random variables.*

2.4. Independence. Two events $A_1, A_2 \subseteq \Omega$ are called **independent** if $\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1)\mathbb{P}(A_2)$ (and they are called **dependent** otherwise). In words, two events are independent if the probability of both occurring at the same time is just the product of the probabilities of the separate events.

This definition extends to more than two events. A set of events $\mathcal{A}$ is called independent if for all subsets $\mathcal{A}' \subseteq \mathcal{A}$,

$$\mathbb{P}\left(\bigcap_{A \in \mathcal{A}'} A\right) = \prod_{A \in \mathcal{A}'} \mathbb{P}(A).$$

It is also possible to define independence for random variables. Two random variables $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent, if and only if

$$\mathbb{P}(\mathbf{X}_1 \leq x_1 \text{ and } \mathbf{X}_2 \leq x_2) = \mathbb{P}(\mathbf{X}_1 \leq x_1)\,\mathbb{P}(\mathbf{X}_2 \leq \mathbf{x}_2).$$

for all x_1 and x_2.

[1] If we are working with a general sigma-algebra, $\mathbf{X}$ is required to be a measureable function.

2.5. Conditional probability. The **conditional probability** of A_1 given A_2 is defined as

$$\mathbb{P}\left(A_1 \mid A_2\right) = \frac{\mathbb{P}\left(A_1 \cap A_2\right)}{\mathbb{P}\left(A_2\right)}.$$

Note that this definition only makes sense if $\mathbb{P}\left(A_2\right) > 0$. The expression $\mathbb{P}\left(A_1 \mid A_2\right)$ is interpreted as the probability that the outcome of the experiment is in A_1, given that we already know that it is in A_2.

Note that if A_1 and A_2 are independent, then $\mathbb{P}\left(A_1 \mid A_2\right) = \mathbb{P}\left(A_1\right)$.

3. Expected value

The expectation of a discrete random variable $\mathbf{X}$ with density function p is defined as

$$\mathbb{E}\left[\mathbf{X}\right] = \sum_x xp(x).$$

Suppose that $\mathbf{X}$ and $\mathbf{Y}$ are random variables, and that λ is a real number, then

$$\mathbb{E}\left[\mathbf{X}+\mathbf{Y}\right] = \mathbb{E}\left[\mathbf{X}\right] + \mathbb{E}\left[\mathbf{Y}\right], \qquad \text{and} \qquad \mathbb{E}\left[\lambda X\right] = \lambda\mathbb{E}\left[\mathbf{X}\right].$$

This is referred to as **linearity of expectation**. Shockingly, it holds for any pair of random variables, regardless of their dependence.

Note that expectation does not behave as nicely as one would hope for general functions of random variables, for example, in general we do not have $\mathbb{E}\left[\mathbf{XY}\right] = \mathbb{E}\left[\mathbf{X}\right]\mathbb{E}\left[\mathbf{Y}\right]$, except when $\mathbf{X}$ and $\mathbf{Y}$ are independent.

EXAMPLE. *Let* $\mathbf{X}$ *be a random variable that takes either of the values* $1, -1$*, each with probability* $1/2$*, and suppose that* $\mathbf{Y} = \mathbf{X}$ *(so, obviously,* $\mathbf{X}$ *and* $\mathbf{Y}$ *are dependent). We compute* $\mathbb{E}\left[\mathbf{X}\right] = \mathbb{E}\left[\mathbf{Y}\right] = 0$.

The random variable $\mathbf{X}+\mathbf{Y}$ *has expectation 0 as well, thus confirming linearity of expectation in this case. On the other hand,* $\mathbf{XY} = \mathbf{X}^2$*, which attains the value 1 with probability 1, so* $\mathbb{E}\left[\mathbf{XY}\right] = 1$.

4. Variance

Suppose that $\mathbf{X}$ is a random variable with expected value $\mu = \mathbb{E}\left[\mathbf{X}\right]$. Then its **variance** is defined as

$$\operatorname{Var}\left(\mathbf{X}\right) = \mathbb{E}\left[(\mathbf{X}-\mu)^2\right].$$

Note that the variance of a random variable is always non-negative. It can be used as a crude measure of deviation from the mean. Using linearity of expectation, we find that

$$\operatorname{Var}\left(\mathbf{X}\right) = \mathbb{E}\left[\mathbf{X}^2\right] - \mathbb{E}\left[\mathbf{X}\right]^2,$$

which in many cases provides a more useful expression to compute the variance of a random variable.

5. Some standard distributions

The probabilistic structure of a random variable is called its distribution. There is a number of standard distributions that we will often encounter in this course.

5.1. Binomial distribution. We say that $\mathbf{X}$ follows a **binomial distribution** with parameters n and p (which we will denote by $\mathbf{X} \sim \operatorname{BIN}\left(n, p\right)$) if it takes values in $\{0, 1, \ldots, n\}$ and has density function

$$p(x) = \binom{n}{i} p^i (1-p)^{n-i}.$$

A binomial random variable is interpreted as the number of successes in n (independent) trials, if the success probability is p.

Example. *Suppose that we throw a biased coin (with probability p of heads) n times, then the number of heads observed in this sequence follows a binomial distribution.*

- $\mathbb{E}[\mathbf{X}] = np$;
- $\mathrm{Var}(\mathbf{X}) = np(1-p)$.

5.2. Geometric distribution. We say that $\mathbf{X}$ follows a **geometric distribution** with parameter p (which we will denote by $\mathbf{X} \sim \mathrm{GEO}(p)$) if it takes values in $\{1, 2, \ldots\}$ and has density function

$$p(x) = p(1-p)^{x-1}.$$

A binomial random variable counts the number of (independent) trials until the first success occurs, when the probability of success is p.

Example. *Suppose that we toss a biased coin (with probability p of heads) until heads comes up for the first time, then the number of trials follows a geometric distribution.*

- $\mathbb{E}[\mathbf{X}] = \frac{1}{p}$;
- $\mathbb{E}[\mathbf{X}] = \frac{1-p}{p^2}$.

Remark. *The random variable $\mathbf{X} - 1$ (i.e. the number of failures before the first success) is referred to as a geometric random variable as well. In that case, the geometric distribution has density function $p(x) = p(1-p)^x$, $x = 0, 1, \ldots$, and expectation $\frac{1-p}{p}$. (Note that $Var(\mathbf{X} - 1) = Var(\mathbf{X})$.)*

You should always specify which of the two interpretations of "geometric random variable" you use, in order to avoid confusion.

Bibliography

[Bil95] Patrick Billingsley. *Probability and measure.* 3rd ed. Wiley, 1995.

[Fel68a] William Feller. *An introduction to probability theory and its applications.* Vol. I. John Wiley, 1968.

[Fel68b] William Feller. *An introduction to probability theory and its applications.* Vol. II. John Wiley, 1968.