

NMAI059 – Probability and Statistics 1

Mykhaylo Tyomkyn

Lecture 1 - Introduction. Axioms of Probability

Probability is a term often used in everyday life, in an attempt to measure randomness and uncertainty. But what do we actually mean when we say “The probability of [...] happening is large/small”? We refer to any scenario involving randomness as a *random experiment*, no matter if it is man made or inherent in the nature.

Philosophically, the two most commonly used interpretations of probability are as follows.

- **Frequentist:** How often will the event happen if the experiment is repeated many times?
- **Subjective/Bayesian:** How certain I am about the event to occur in the experiment? What betting odds am I willing to take for it?

There is a lot to be said about perceived advantages and drawbacks of either approach, but ultimately it does not matter to us, as we aim to study Probability as a formal mathematical concept.

As the name of the course stipulates, we shall cover basic concepts of *Probability theory* and *Statistics*. Both study randomness mathematically. So, what is the difference between the two? When dealing with a random experiment, Probability (theory) assumes the underlying ‘rules of randomness’ as given, while Statistics tries to infer them from the experiment itself.

Example 1 *A typical question in Probability: A fair coin is tossed 100 times. How likely are we to see at least 60 ‘heads’?*

Its counterpart in Statistics: An unknown coin is tossed 100 times, and comes up ‘heads’ 60 times. How likely is it that the coin is fair?

In order to do Statistics, inferring unknown probabilities, we need first to develop a good understanding of probabilities when the rules of randomness are given. Hence, we will begin with Probability theory, and move on to Statistics in the second half of the course.

To define probability formally, let us first consider the classical approach that some of you may have seen in high school:

$$\text{Probability} = \frac{\#\text{Relevant outcomes}}{\#\text{All outcomes}}. \quad (1)$$

This works well in many scenarios, such as a single die throw, but has some shortcomings.

1. Quite often the *elementary outcomes* of an experiment are not equally likely, think of a biased coin or a loaded die. We want to have more flexibility in our modelling.

2. We want to extend the notion of probability to experiments ('sample spaces') with infinitely, or even uncountably many outcomes.

Example 2 (Bertrand's paradox) *Given a circle of radius 1, what is the probability that a randomly chosen chord forms a central angle of at least 120 degrees? There are three different methods of computing said probability, which all appear very logical, yet yield different results: 1/3, 1/4, 1/2.*

It turns out that the classical probability notion, and our intuition based thereon, do not apply in infinite, let alone uncountable, sample spaces. The workaround: let the probability be "what we define it to be", as long as it follows certain natural rules/axioms.

Definition 1 (Probability space) *A probability space is a triple $(\Omega, \mathcal{F}, \mathbb{P})$, where*

- Ω is a set, called the **sample space**.
- $\mathcal{F} \subseteq 2^\Omega$ is the **event space**,¹ satisfying the σ -algebra axioms (see below).
- $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$ is a function, called the **probability function** (also "probability measure", "probability distribution"), satisfying Kolmogoroff's axioms (see below).

The interpretation of the above is as follows. Each $\omega \in \Omega$ is an *elementary outcome* of a random experiment. For instance, for a single die throw we can take $\Omega = \{1, 2, 3, 4, 5, 6\}$.

\mathcal{F} is the set of all "events", i.e., subsets $A \subseteq \Omega$ about which we make probability statements. We posit that \mathcal{F} is a σ -algebra.

Definition 2 (σ -algebra) *Let Ω be a set. A set $\mathcal{F} \subseteq 2^\Omega$ is a σ -algebra ("sigma algebra") if*

- (i) $\emptyset \in \mathcal{F}$ and $\Omega \in \mathcal{F}$.
- (ii) $A \in \mathcal{F} \Rightarrow \Omega \setminus A \in \mathcal{F}$.
- (iii) $A_1, A_2, \dots \in \mathcal{F} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$. *That is, countable unions of sets in \mathcal{F} are again in \mathcal{F} . Note that by (ii) and De Morgan's law, countable intersections are also in \mathcal{F} .*

When Ω is finite or countably infinite, we always take $\mathcal{F} = 2^\Omega$, i.e. we may speak of probability of any collection of outcomes. In this case, Definition 2 becomes redundant (there is a reason why we take a more restrictive approach with uncountable sample spaces, more on this further down).

Finally, \mathbb{P} assigns to the events their probabilities. We demand that it satisfies the following axioms.

Definition 3 (Kolmogoroff's axioms) *The probability function \mathbb{P} must satisfy*

- (i) $\mathbb{P}(\emptyset) = 0$ and $\mathbb{P}(\Omega) = 1$.
- (ii-) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ for any disjoint events A, B .
- (ii) *More generally, $\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$ for any sequence² of pairwise disjoint³ events A_1, A_2, \dots*

¹ 2^Ω denotes the power set of Ω . That is, $2^\Omega = \{A \subseteq \Omega\}$.

²Note that the value of the sum does not depend on the summation order, as all summands are non-negative. You may recall the notion of "absolutely convergent series" from Analysis I.

³Meaning $A_i \cap A_j = \emptyset$ for any $i \neq j$.

The last item in this list is usually referred as σ -**additivity** (sigma-additivity).

When Ω is finite or countably infinite, as we use $\mathcal{F} = 2^\Omega$, we can work with the following simplified definition.

Definition 4 (Countable probability spaces) *A countable probability space is a tuple (Ω, \mathbb{P}) comprising a countable set Ω and, with $\mathcal{F} = 2^\Omega$, a function $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$, satisfying*

- $\mathbb{P}(\emptyset) = 0, \mathbb{P}(\Omega) = 1,$
- $\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\{\omega\}),$ for every $A \in \mathcal{F}.$

In particular, we must have $\sum_{\omega \in \Omega} \mathbb{P}(\{\omega\}) = 1.$ Conversely, every sequence of non-negative reals whose total sum is 1 gives rise to a probability distribution. Let us now look at some basic examples.

Example 3 *Let Ω be finite, and for all $\omega \in \Omega$ set $\mathbb{P}(\{\omega\}) = 1/|\Omega|.$ Then, for any $A \in \mathcal{F}$ we have*

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|}.$$

*This is the **uniform measure/distribution** and corresponds to the classical approach described informally in (1).*

Remarkably, when Ω is countably infinite, say $\Omega = \mathbb{N},$ a uniform probability measure does not exist! This is because there exists no constant function $f : \mathbb{N} \rightarrow [0, 1]$ satisfying $\sum_{i=1}^{\infty} f(i) = 1.$ Thus, there is no such thing as a ‘(uniformly) random integer’.

Example 4 *Let $\Omega = \mathbb{N} = \{1, 2, \dots\}$ and $\mathbb{P}(\{i\}) = 2^{-i}$ for all $i.$ Then*

$$\sum_{\omega \in \Omega} \mathbb{P}(\{\omega\}) = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \dots = 1.$$

So, this gives rise to a probability distribution on \mathbb{N} via⁴

$$\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\{\omega\}) = \sum_{i \in A} 2^{-i}.$$

*This is an instance of the **geometric distribution**, which we will define later more generally.*

Exercise 1 *In the above probability model, what is the probability that the randomly chosen number is even? In other words, find $\mathbb{P}(A),$ where $A = \{2, 4, \dots\}$ is the set of positive even numbers.*

When Ω is uncountable, for instance $\Omega = \mathbb{R}$ or an interval in $\mathbb{R},$ one needs to be more careful about the event space $\mathcal{F},$ which is why we bothered defining a σ -algebra. Allowing $\mathcal{F} = 2^\Omega$ as in the countable case would lead to counterintuitive phenomena or even contradictions.

For example, suppose $\Omega = [0, 1], \mathcal{F} = 2^\Omega,$ and we want to define the uniform measure $\mathbb{P} : \mathcal{F} \rightarrow [0, 1].$ It is natural to require that \mathbb{P} satisfy $\mathbb{P}([a, b]) = b - a$ for all $0 \leq a \leq b \leq 1,$ and that \mathbb{P} be translation invariant, i.e., shifting A by a constant would not change its probability (as long as the image remains inside $[0, 1]$). Surprisingly, it turns out that these natural demands are impossible to reconcile with σ -additivity (look up “Vitali set”).

Having said that, these phenomena can be ignored for most practical purposes.

Let us now establish some fundamental properties of probability spaces.

⁴Technically on $2^\mathbb{N},$ but this is a commonly used shorthand

Theorem 1 Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we have

1. $\mathbb{P}(A) + \mathbb{P}(\Omega \setminus A) = 1$ for all $A \in \mathcal{F}$.
2. $A \subseteq B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$ for all $A, B \in \mathcal{F}$. ('monotonicity')
3. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$ for all $A, B \in \mathcal{F}$. ('inclusion-exclusion')
4. $\mathbb{P}(\bigcup_{n=1}^{\infty} A_n) = \lim_{n \rightarrow \infty} \mathbb{P}(A_n)$ for any $A_1, A_2, \dots \in \mathcal{F}$ with $A_1 \subseteq A_2 \subseteq A_3 \subseteq \dots$. ('continuity')
5. $\mathbb{P}(\bigcup_{n=1}^{\infty} A_n) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_n)$ for any $A_1, A_2, \dots \in \mathcal{F}$. ('union bound')

Note that the σ -additivity axiom holds with $=$ but requires the events to be pairwise disjoint, while the union bound holds 'merely' with \leq , but for any sequence of events.

Proof Statements 1., 2. and 3. follow immediately from the axioms.

To prove 4. first note that the limit always exists, as the sequence $(\mathbb{P}(A_n))_n$ is increasing (by statement 2.) and bounded since $\mathbb{P}(A_n) \leq 1$ for all n . Consider now the events $B_n = A_n \setminus A_{n-1}$ and $B_1 = A_1$. Note that the events $(B_n)_n$ are pairwise disjoint and $\bigcup_{n=1}^{\infty} B_n = \bigcup_{n=1}^{\infty} A_n$. Therefore, by σ -additivity,

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} B_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(B_n) = \lim_{n \rightarrow \infty} \sum_{i=1}^n \mathbb{P}(B_i) = \lim_{n \rightarrow \infty} \mathbb{P}\left(\bigcup_{i=1}^n B_i\right) = \lim_{n \rightarrow \infty} \mathbb{P}(A_n).$$

To prove 5., first show that it holds for finitely many events, using inclusion-exclusion and induction (exercise). To extend it to infinite sequences, consider the events $C_n = \bigcup_{i=1}^n A_i$. Note that $C_1 \subseteq C_2 \subseteq \dots$ and $\bigcup_{i=1}^{\infty} C_n = \bigcup_{i=1}^{\infty} A_n$. So, we may apply 4. to obtain

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} C_n\right) = \lim_{n \rightarrow \infty} \mathbb{P}(C_n) = \lim_{n \rightarrow \infty} \mathbb{P}\left(\bigcup_{i=1}^n A_i\right).$$

Now, for each n we have

$$\mathbb{P}\left(\bigcup_{i=1}^n A_i\right) \leq \sum_{i=1}^n \mathbb{P}(A_i) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i),$$

and so this also must hold in the limit:

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n \rightarrow \infty} \mathbb{P}\left(\bigcup_{i=1}^n A_i\right) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i) = \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

□