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 subjective advantages and drawbacks of either approach, but ultimately it does not matter for 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 us may have seen in high school:

$$\text{Probability} = \frac{\#\text{Relevant outcomes}}{\#\text{All outcomes}}.$$

This works well in many scenarios (e.g. a single die throw), but has some shortcomings.

1. Oftentimes 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 infinite, or even uncountable numbers of 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 estimating 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 certain axioms (for now, and for the most part of the course we can take* $\mathcal{F} = 2^\Omega$*)*
- $\mathbb{P} : \mathcal{F} \to [0,1]$ *is a function, called the* ***probability function*** *(also "probability measure", "probability distribution"), satisfying Kolmogoroff's axioms (see below).*

The interpretation of the above: each $\omega \in \Omega$ is an *elementary outcome* of a random experiment (e.g., 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. When Ω is finite or countably infinite, we always put $\mathcal{F} = 2^\Omega$, i.e., we may speak of probability of any collection of outcomes (there is a reason why we take a more restrictive approach when dealing 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 2 (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 sequence*[2] *of pairwise disjoint*[3] *events* $A_1, A_2, \ldots,$.

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

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

[2] Note that the value of the sum does not depend on the summation order, as all summands are non-negative ("absolute convergence").

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

- $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\Omega) = 1$,
- $\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\{\omega\})$, *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 model we discussed in the introduction.*

Remarkably, when Ω is countably infinite, say $\Omega = \mathbb{N}$, a uniform measure does not exist! This is because there exists no constant function $f : \mathbb{N} \to [0, 1]$ 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*

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

So, this gives rise to a probability distribution on $\mathbb{N}$ via[4]

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

*This is an instance of the **geometric distribution**, we will define it 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 $\mathcal{F}$. Allowing $\mathcal{F} = 2^{\Omega}$ as before would lead to counterintuitive phenomena or even contradictions.

For example, suppose $\Omega = [0, 1]$, $\mathcal{F} = 2^{\Omega}$, and want to define the uniform measure $\mathbb{P} : \mathcal{F} \to [0, 1]$. It is natural to require that $\mathbb{P}$ satisfy $\mathbb{P}([a, b]) = b - a$ for all $0 \le a \le b \le 1$, 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.

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

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

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. $\mathbb{P}(\bigcup_{n=1}^{\infty} A_n) = \lim_{n\to\infty} \mathbb{P}(A_n)$ *for any* $A_1, A_2, \cdots \in \mathcal{F}$ *with* $A_1 \subseteq A_2 \subseteq A_3 \subseteq \ldots$ *('continuity')*

5. $\mathbb{P}(\bigcup_{n=1}^{\infty} A_n) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_n)$ *for any* $A_1, A_2, \cdots \in \mathcal{F}$. *('union bound')*

Note that the σ-additivity axiom holds with $=$ but requires that the events are 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,

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

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 $C_1 \subseteq C_2 \subseteq \ldots$ and $\bigcup_{i=1}^{\infty} C_n = \bigcup_{i=1}^{\infty} A_n$. So, apply statement 4. we obtain

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

Now, for each n we have

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

and so this also must hold in the limit:

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

□