4. Foundation of probability

Contents

(All variables are real and one-dimensional unless otherwise specified.)

After all, what is **probability**? No one exactly knows; mathematicians, physicists, philosophers, and theologians may give quite different answers. Ultimately, we do not even know if probability is something "real".

1. Kolmogorov's axioms

First, let us look at probability from a mathematician's perspective. **Kolmogorov** defines probability with his famous three **axioms**, which are now often regarded as the cornerstone for a rigorous discussion on probability. The following paragraphs intend to illustrate the axioms in plain words while keeping their ideas intact.

1.1 Premises

- Ω is a **sample space** containing the possible outcomes of whatever you are interested in. For example, the possible sample space of rolling a dice once is $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- E is an **event** that you can observe. It is formally a subset of Ω. In terms of the dice, $E = \{1, 2, 3\}$ means "I observe one, two, or three", $E = \emptyset$ means "I observe nothing", and $E = \Omega$ means "I observe one, two, three, four, five, or six".
- F is an **event space**, i.e. a set of the possible events you can observe. If $E \in F$, a valid F must contain the event's complement $\bar{E} \equiv \Omega \setminus E$ and its union $E \cup E'$ with some other event $E' \in F$.
- P is a **function** that maps a **set** to a **real number**. Conventionally we call this **probability.**

1.2 The three axioms

We now have the ingredients to state the three axioms, viz. **non-negativity**, **unitarity**, and **countable additivity**.

- **Non-negativity.**

$$P(E) \geq 0 \quad \forall E \in F$$

- **Unitarity.**

$$P(\Omega) = 1$$

- **Countable additivity.** For an **infinite** set of **mutually exclusive** events $\{E_1, E_2, \ldots\}$, i.e. $E_i \cap E_j \equiv \emptyset$ for all $i \neq j$,

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)\,.$$

1.3 Properties of probability

With these three axioms, we can derive several properties of probability.

- **Monotocity.**

$$E \subseteq E' \Rightarrow P(E) \leq P(E')$$

- **The complement rule.**

$$P(\bar{E}) = 1 - P(E)$$

- **Probability of a null event.**

$$P(\emptyset) = 0$$

- **Range of probability.**

$$0 \leq P(E) \leq 1$$

- **Sum rule.**

$$P(E \cup E') = P(E) + P(E') - P(E \cap E')$$

- **Finite additivity.** For a **finite** set of mutually exclusive events $\{E_1, E_2, \ldots, E_n\}$,

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i)\,.$$

1.4 Example: infinitely wide uniform distribution

Can we have an infinitely wide uniform distribution $\mathcal{U}_x(-\infty, +\infty)$? Because a uniform distribution over a finite range (a, b) is

$$\mathcal{U}_x(a,b) = \begin{cases} \dfrac{1}{b-a} & (a<x<b) \\ 0 & (\text{otherwise}) \end{cases},$$

$\mathcal{U}_x(-\infty,+\infty)$ intuitively gives us zero for all x. What does this mean?

Solution. Kolmogorov's axioms **prohibits** the existence of an "infinitely wide uniform distribution". We can prove it by contradiction. Assume $\mathcal{U}_x(-\infty,+\infty)$ exists, then let E_i be the event of observing $x \in [n, n+1)$, where $n = (-1)^i \lfloor i/2 \rfloor$. As $x \in (-\infty,+\infty)$, each observed x corresponds to **one and only one event** from an infinite set of mutually exclusive events $\{E_1, E_2, \ldots\}$, so the sample space is $\Omega = \bigcup_{i=1}^{\infty} E_i$. Because the distribution is uniform, the probability of E_i would be some constant p for all i, while the first axiom requires $p \geq 0$. Then according to the third axiom,

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i) = p + p + p + \cdots .$$

The sum diverges unless $p = 0$, so we must let $p = 0$ and $P\left(\bigcup_{i=1}^{\infty} E_i\right) = 0$. However, the choice consequently contradicts with the second axiom $P(\Omega) = P\left(\bigcup_{i=1}^{\infty} E_i\right) \equiv 1$. Since it violates either the second or the third axiom under the first axiom, the distribution cannot exist.

2. Nature of probability

However rigorous Kolmogorov's axioms are, they do not tell us the **meaning** of probability. This is a matter of **philosophy**. Regarding this question, there are three major schools, viz. the classical school, the **frequentist** (aka ontic) school, and the **Bayesian** (aka epistemic) school.

Before proceeding, first think about a question: What is the probability for a 180cm-tall Englishman's left eye to be green in colour? How do you assess its value?

2.1 The classical school

It is chiefly founded on Laplace's **principle of indifference**. Suppose an experiment has N **possible outcomes**. If N_A outcomes are regarded as the occurrence of an event A, the probability of A is

$$P(A) = \frac{N_A}{N}.$$

For example, because a coin only has two faces, the probability to get a head from tossing a coin equals 0.5, so does that to get a tail. This approach has two obvious drawbacks.

- Why must all the outcomes be **equiprobable**? Not all coins are fair.
- What happens as $N \to \infty$? There are **infinitely many possible** outcomes for a continuous random variable. How is its probability defined then?

After all, we are not even sure about how many possible outcomes there are: why can't a coin, as Stephen Chow's *Shaolin Soccer* makes fun of, stand up on its side after a toss? The classical school has therefore basically been displaced by other modern schools and perhaps used only for introducing probability to novices.

2.2 The frequentist school

It regards probability as **relative frequency**: the probability of an experiment's outcome A is

$$P(A) = \lim_{n \to \infty} \frac{n_A}{n},$$

where A is observed for n_A times after **repeating** the experiment for n times. This definition implies that probability belongs to an experiment and is its intrinsic property. For example, if a coin's probability of head is 0.4, you get a head in 40% of your tosses. This definition looks objective and scientific, but it fails in two aspects.

- How is probability defined for experiments that **cannot be repeated** and events that happens **once** only? What is the probability for Hong Kong to have an earthquake? What is the probability for you to die in a traffic accident?
- In a repeatable experiment, I observe A for n_A times in n trials, while you observe it for m_A times in m trials. Is my frequency $\frac{n_A}{n}$ or your $\frac{m_A}{m}$ more accurate? Or should we combine them as something like $\frac{n_A + m_A}{n + m}$? After all, is probability still **constant** or in fact **constantly changing**?

Frequentists argue that we can never know an event's **exact** probability, but we can only **estimate** its true value with experiments. Hence, probability is measured probabilistically—What is the probability that our frequency matches the true probability?—Wait, what?

2.3 The Bayesian school

It interprets probability as **degree of belief** with Bayes' theorem. Recall Bayes' theorem:

$$P(A \mid B) = \frac{P(B \mid A)}{\sum_{A'} P(B \mid A')P(A')} \times P(A)\,.$$

The Bayesian school argues that $P(A)$ on the right hand side measures one's **prior** belief in an event A, whereas $P(A \mid B)$ on the left hand sides measures his **posterior** belief in the same event after obtaining extra information B. The person is said to have **updated** his belief.

For example, if I tell you a coin is fair, you believe that it is equiprobable to get a head and a tail, so its probability of head is 0.5; once I tell you the coin is completely biased to its head, you will update your belief and believe that the probability of head is 1 instead. You judge merely according to what you are told; you have not even touched the coin. In other words, probability of an event does not really belong to the event itself but to its **observer**.

Consequently, the major criticism against the Bayesian school challenges its **subjectivity**.

- How is the **first** prior degree of belief defined?
- Why does probability depend on who is observing? Why would a coin's probability of head ever change simply due to some irrelevant people's bluffing?

Some Bayesians rebut with the so-called **objective Bayesian interpretation**. They advocate that probability is "objectively subjective" and thus observer-invariant: it measures the **common** degree of belief of **any** sensible person facing the given knowledge. Finally, remember: using Bayes' theorem does not automatically make one Bayesian; **over-interpretation** does.

If you are interested, read **Probability Interpretations (https://en.wikipedia.org/wiki/Probability_interpretations)** on Wikipedia. Now, return to the probability for a 180cm-tall Englishman to have a green left eye: will you change your answer given that his right eye is green in colour? Why or why not?

3. Famous paradoxes

Finally, let us confront ourselves with some famous paradoxes to deepen our understanding in probability, especially in how frequestists and Bayesians interpret the same number distinctly.

3.1 Bertrand's boxes (https://en.wikipedia.org/wiki/Bertrand%27s_box_paradox)

There are three boxes that look identical. However, one contains two gold coins, one contains two silver coins, and one contains one gold coin and one silver coin. You randomly choose a box and draw one coin: it is gold. What is the probability that your box contains two gold coins?

One half? No, the answer is two thirds.

Solution. Denote the three boxes with GG, SS, and GS. The boxes look the same, so it is equiprobable to choose any one of them.

$$P(\mathrm{GG}) = P(\mathrm{SS}) = P(\mathrm{GS}) = \frac{1}{3}$$

If your box is GG, you always draw a gold coin G. If it is SS, never. If it is GS, half of the time.

$$\begin{cases} P(\mathrm{G} \mid \mathrm{GG}) &= 1 \\ P(\mathrm{G} \mid \mathrm{SS}) &= 0 \\ P(\mathrm{G} \mid \mathrm{GS}) &= 1/2 \end{cases}$$

Now put everything into Bayes' theorem to get the final answer.

$$P(\mathrm{GG} \mid \mathrm{G}) = \frac{P(\mathrm{G} \mid \mathrm{GG})P(\mathrm{GG})}{P(\mathrm{G} \mid \mathrm{GG})P(\mathrm{GG}) + P(\mathrm{G} \mid \mathrm{SS})P(\mathrm{SS}) + P(\mathrm{G} \mid \mathrm{GS})P(\mathrm{GS})} = \frac{2}{3}$$

Frequentist interpretation. If I keep choosing a box and then drawing a coin, around $2/3$ of the boxes from which I draw a gold coin contains two gold coins.

Bayesian interpretation. If I choose a box and know nothing about it, my degree of belief in "the box contains two gold coins" is $1/3$. After knowing that it has at least one gold coin, my degree of belief rises to $2/3$.

3.2 The Monty Hall problem (https://en.wikipedia.org/wiki/Monty_Hall_problem)

Monty Hall is a Canadian-American host, and the Monty Hall problem originates from his game show *Let's Make A Deal*. His problem is formally similar to Bertrand's boxes, but it **confuses so many people, including great mathematicians,** that it has perhaps become the most famous paradox about probability.

Monty Hall shows you three identical doors, viz. door 1, door 2, and door 3. There is a prize behind one of the doors and nothing behind the others. You can choose one door and win the thing behind it. You choose, say, door 1. Monty Hall knows where the prize is and wants to tease you, so he opens a door behind which there is nothing; suppose that he opens door 3. Then he lets you choose your door again. Does a switch to, in this case, door 2 increase the chance of winning the prize?

The answer is yes.

Solution. Let u_i be "you choose door i" and π_i be "the prize is behind door i". After you have chosen door 1, the prize can be behind any door, so

$$P(\pi_1 \mid u_1) = P(\pi_2 \mid u_1) = P(\pi_3 \mid u_1) = \frac{1}{3}\,.$$

Intuitively—as many scholars think—once Monty Hall lets us know π_3 is impossible, $P(\pi_3 \mid u_1)$ is eliminated, leaving us

$$P(\pi_1 \mid u_1) = P(\pi_2 \mid u_1) = \frac{1}{2},$$

so switching to door 2 does not seem superior to staying at door 1. However, this logic **fails** because it has neglected **Monty Hall's possibility** to open door 2 instead of door 3. He is not forced to open door 3; instead, he only needs to open a door that **you do not choose** and **has nothing behind it**. If the prize is behind door 2, he indeed must open door 3. In contrast, he may open either door 2 or door 3 if the prize is in fact behind door 1. Considering this, let m_i be "Monty Hall opens door i".

$$\begin{cases} P(m_3 \mid u_1, \pi_1) &= 1/2 \\ P(m_3 \mid u_1, \pi_2) &= 1 \\ P(m_3 \mid u_1, \pi_3) &= 0 \end{cases}$$

Then Bayes' theorem gives

$$P(\pi_2 \mid u_1, m_3) = \frac{P(m_3 \mid u_1, \pi_2) P(u_1, \pi_2)}{\sum_{i=1}^3 P(m_3 \mid u_1, \pi_i) P(u_1, \pi_i)} = \frac{2}{3},$$

where $P(u_1, \pi_i) = P(u_1)P(\pi_i)$ and $P(\pi_1) = P(\pi_2) = P(\pi_3) = 1/3$. As a result, switching to door 2 is more advantageous than staying at door 1.

Frequentist interpretation. Suppose I keep playing this game with Monty Hall. I can win the prize in around $2/3$ of the occasions that I change, while I can win the prize in around $1/3$ of the occasions that I do not change.

Bayesian interpretation. If I choose a door and do not know anything about it, my degree of belief in "the prize is behind my door" is $1/3$. After knowing that there is nothing behind a door that I do not choose, my degree of belief in "the prize is behind the other closed door" rises to $2/3$, whereas that in "the prize is behind my door" is still $1/3$.

3.3 The boy-or-girl paradox (https://en.wikipedia.org/wiki/Boy_or_Girl_paradox) (https://en.wikipedia.org/wiki/Boy_or_Girl_paradox)

Mr Smith has two children. Given that at least one child is a boy, how likely does Mr Smith have two boys? Assume that a child is equiprobably either a boy or a girl and that the siblings' sexes are independent.

It turns out that the question is **ambiguous**. Its answer depends on the **exact experimental procedures**, i.e. how the fact "at least one child is a boy" is obtained.

Solution. Mr Smith may have two boys BB, two girls GG, an older boy with a younger girl BG, and an older girl with a younger boy GB. The four scenarios are assumed equiprobable.

$$P(\mathrm{BB}) = P(\mathrm{GG}) = P(\mathrm{BG}) = P(\mathrm{GB}) = \frac{1}{4}$$

Three scenarios BB, BG, and GB match "at least one child is a boy", denoted as B. The probability of BB conditional on B is hence

$$P_1(\mathrm{BB} \mid \mathrm{B}) = P_1(\mathrm{B} \mid \mathrm{BB}) \times \frac{P(\mathrm{BB})}{P_1(\mathrm{B})} = 1 \times \frac{1/4}{3/4} = \frac{1}{3},$$

where $P_1(\mathrm{B})$ can be expanded as

$$\begin{aligned} P_1(\mathrm{B}) &= P_1(\mathrm{B} \mid \mathrm{BB})P(\mathrm{BB}) + P_1(\mathrm{B} \mid \mathrm{GG})P(\mathrm{GG}) \\ &\quad + P_1(\mathrm{B} \mid \mathrm{BG})P(\mathrm{BG}) + P_1(\mathrm{B} \mid \mathrm{GB})P(\mathrm{GB}) \\ &= 1 \times \frac{1}{4} + 0 \times \frac{1}{4} + 1 \times \frac{1}{4} + 1 \times \frac{1}{4} = \frac{3}{4}. \end{aligned}$$

However, if you conclude that at least one of Mr Smith's children is a boy because you have visited Mr Smith and only seen a boy, the conditional probability becomes

$$P_2(\mathrm{BB} \mid \mathrm{B}) = P_2(\mathrm{B} \mid \mathrm{BB}) \times \frac{P(\mathrm{BB})}{P_2(\mathrm{B})} = 1 \times \frac{1/4}{1/2} = \frac{1}{2}$$

with

$$\begin{aligned} P_2(\mathrm{B}) &= P_2(\mathrm{B} \mid \mathrm{BB})P(\mathrm{BB}) + P_2(\mathrm{B} \mid \mathrm{GG})P(\mathrm{GG}) \\ &\quad + P_2(\mathrm{B} \mid \mathrm{BG})P(\mathrm{BG}) + P_2(\mathrm{B} \mid \mathrm{GB})P(\mathrm{GB}) \\ &= 1 \times \frac{1}{4} + 0 \times \frac{1}{4} + \frac{1}{2} \times \frac{1}{4} + \frac{1}{2} \times \frac{1}{4} = \frac{1}{2}. \end{aligned}$$

The apparent inconsistency arises due to the ambiguous meaning of $P(\mathrm{B} \mid \mathrm{BG})$ [and $P(\mathrm{B} \mid \mathrm{GB})$, which follows a parallel argument]. In the former case, we actually **know both children's sexes**, but in the latter case, we only **know one child's sex**, which happens to be a boy.

$$P_1(\mathrm{B} \mid \mathrm{BG}) = P(\mathrm{BG\ has\ at\ least\ one\ boy}) = 1$$

$$P_2(\mathrm{B} \mid \mathrm{BG}) = P(\mathrm{the\ child\ is\ a\ boy} \mid \mathrm{see\ a\ child\ from\ BG}) = 1/2$$

Variant. A [more mind-blowing variant (https://en.wikipedia.org/wiki/Boy_or_Girl_paradox#Information_about_the_child)](https://en.wikipedia.org/wiki/Boy_or_Girl_paradox#Information_about_the_child) of this paradox adds an extra condition: at least one of Mr Smith's child is a boy born on Tuesday. This piece of seemingly **irrelevant information** turns out to be able to change the conditional probability drastically. If we see a boy who happens to be born on Tuesday, the conditional probability is still $1/2$ (following the analysis of P_2). In contrast, if we actually know both children's sexes and weekdays of birth, the conditional probability rises from $1/3 \approx 0.33$ to around 0.48 (following the analysis of P_1), assuming equal probability of a child's weekday of birth. In general, if it is given

that at least one boy possesses a feature that occurs with a probability p, the conditional probability of P_1 is $\frac{2-p}{4-p}$, whereas that of P_2 remains $1/2$.

3.4 Bertrand's circle (https://en.wikipedia.org/wiki/Bertrand_paradox_(probability)) (https://en.wikipedia.org/wiki/Bertrand_paradox_(probability))

It asks for the probability for a **random** chord in a unit circle to be longer than three units. It is morally similar to the boy-and-girl paradox: the answer depends on **exact experimental procedures**, i.e. how a random chord is generated.