Probability distributions

1 Discrete distributions

When we toss a coin, we have no idea whether it will land heads or tails. However, there is a different sense in which the behaviour of the coin is highly predictable: if it is tossed many times, then the proportion of heads is very likely to be close to 1/2.

In order to study this phenomenon mathematically, we need to model it, and this is done by defining a *sample space*, which represents the set of possible outcomes, and a *probability distribution* on that space, which tells you their probabilities. In the case of a coin, the natural sample space is the set $\{H, T\}$, and the obvious distribution assigns the number 1/2 to each element. Alternatively, since we are interested in the number of heads, we could use the set $\{0, 1\}$ instead: after one toss, there is a probability of 1/2 that the number of heads is 0 and a probability of 1/2 that it is 1. More generally, a (discrete) sample space is simply a set Ω, and a probability distribution on Ω is a way of assigning a non-negative real number to each element of Ω, such that the sum of all these numbers is 1. The number assigned to a particular element of Ω is then interpreted as the probability that that some corresponding outcome will occur, the total probability being 1.

If Ω is a set of size n, then the *uniform distribution* on Ω is the probability distribution that assigns a probability of $1/n$ to each element of Ω. However, it is often more appropriate to assign different probabilities to different outcomes. For example, given any real number p between 0 and 1, the *Bernoulli distribution with parameter* p on the set $\{0, 1\}$ is the distribution that assigns the number p to 1 and $1 - p$ to 0. This can be used to model the toss of a biased coin.

Suppose now that we toss an unbiased coin n times. If we are interested in the outcome of every toss, then we would choose the sample space consisting of all possible sequences of 0s and 1s of length n. For instance, if $n = 5$, a typical element of the sample space is 01101. (This particular element represents the outcome tails, heads, heads, tails, heads, in that order.) Since there are 2^n such sequences and they are all equally likely, the appropriate distribution on this space will be the uniform one, which assigns a probability of $1/2^n$ to each sequence.

But what if we are interested not in the particular sequence of heads and tails but just in the *total number of heads*? In that case, we could take as our sample space the set $\{0, 1, 2, \ldots, n\}$. The probability that the total number of heads is k is 2^{-n} times the number of sequences of 0s and 1s that contain exactly k 1s. The latter is $\binom{n}{k} = \frac{n!}{k!(n-k)!}$, so the probability we assign to k is $p_k = \binom{n}{k}2^{-n}$.

More generally, for a sequence of n independent experiments, each with the same probability p of success, the probability of a given sequence of k successes and $n - k$ failures is $p^k(1 - p)^{n-k}$. So, the probability of having exactly k successes is $p_k = \binom{n}{k}p^k(1 - p)^{n-k}$. This is called the *binomial distribution* with parameters n and p. It models the number of heads if you toss a biased coin n times, for example.

Suppose we perform such experiments for as long as we need to in order to obtain one success. When k experiments are performed, the probability of getting $k - 1$ failures followed by a success is $p_k = (1 - p)^{k-1}p$. Therefore, this formula gives us the distribution of the number of experiments up to the first success. It is called the *geometric distribution* of parameter p. In particular, the number of tosses of a fair coin needed to get the first head is geometric of parameter 1/2. Notice that our sample space is now the set of all non-negative integers—in particular, it is infinite. So in this case the condition that the probabilities add up to 1 is that a certain infinite series (the series $\sum_{k=1}^{\infty} p_k$) converges to 1.

Now let us imagine a somewhat more complicated experiment. Suppose we have a radioactive source that occasionally emits an alpha particle. It is often reasonable to suppose that these emissions are independent, and equally likely to occur at any time. If the average number of emissions per minute is λ, say, then what is the probability that during any given minute there will be k particles emitted?

One way to think about this question is to divide up the minute into n equal intervals, for some large n. If n is large enough, then the probability of two emissions occurring in the same interval is

so small that it can be ignored, and therefore, since the average number of emissions per minute is λ, the probability of an emission during any given interval must be approximately λ/n. Let us call this number p. Since the emissions are independent, we can now regard the number of emissions as the number of successes when we do n trials, each with probability p of success. That is, we have the binomial distribution with parameters n and p, where $p = \lambda/n$.

Notice that as n gets larger, p gets smaller. Also, the approximations just made become better and better. It is therefore natural to let n tend to infinity and study the resulting "limiting distribution." It can be checked that, in the limit as $n \to \infty$, the binomial probabilities converge to $p_k = e^{-\lambda}\lambda^k/k!$. These numbers define a distribution on the set of all non-negative integers, known as the *Poisson distribution* of parameter λ.

2 Probability spaces

Suppose that I throw a dart at a dartboard. Not being very good at darts, I am not able to say very much about where the dart will land, but I can at least try to model it probabilistically. The obvious sample space to take consists of a circular disk, the points of which represent where the dart lands. However, now there is a problem: if I look at any particular point in the disk, the probability that the dart will land at precisely that point is zero. So how do I define a probability distribution?

A clue to the answer lies in the fact that it seems to be perfectly easy to make sense of a question such as "What is the probability that I will hit the bull's eye?" In order to hit the bull's eye, the dart has to land in a certain region of the board, and the probability of this happening does not have to be zero. It might, for instance, be equal to the area of the bull's eye region divided by the total area of the board.

What we have just observed is that even if we cannot assign probabilities to individual *points* in the sample space, we can still hope to give probabilities to *subsets*. That is, if Ω is a sample space and A is a subset of Ω, we can try to assign a number $\mathbb{P}(A)$ between 0 and 1 to the set A. This represents the probability that the random outcome belongs to the set A, and can be thought of as something like a notion of "mass" for the set A.

For this to work, we need $\mathbb{P}(\Omega)$ to be 1 (since the probability of getting *something* in the sample space must be 1). Also, if A and B are disjoint subsets of Ω, then $\mathbb{P}(A \cup B)$ should be $\mathbb{P}(A)+\mathbb{P}(B)$. From this it follows that if $A_1, \ldots, A_n$ are all disjoint, then $\mathbb{P}(A_1\cup, \ldots, \cup A_n)$ is equal to $\mathbb{P}(A_1) + \cdots + \mathbb{P}(A_n)$. Actually, it turns out to be important that this should be true not just for finite unions but even for COUNTABLY INFINITE ones as well. (Related to this point is the fact that one does not attempt to define $\mathbb{P}(A)$ for *every* subset A of Ω but just for MEASURABLE SUBSETS. For our purposes, it is sufficient to regard $\mathbb{P}(A)$ as defined whenever A is a set we can actually define.)

A *probability space* is a sample space Ω together with a function $\mathbb{P}$, defined on all "sensible" subsets A of Ω, that satisfies the conditions mentioned in the previous two paragraphs. The function $\mathbb{P}$ itself is known as a *probability measure* or *probability distribution*. The term *probability distribution* is often preferred when we specify $\mathbb{P}$ concretely.

3 Continuous probability distributions

There are three particularly important distributions defined on subsets of $\mathbb{R}$, of which two will be discussed in this section. The first is the *uniform distribution* on the interval $[0, 1]$. We would like to capture the idea that "all points in $[0, 1]$ are equally likely." In view of the problems mentioned above, how should we do this?

A good way is to take seriously the "mass" metaphor. Although we cannot calculate the mass of an object by adding up the masses of all the infinitely small points that make up the object, we can assign to those points a *density*, and integrate it. That is exactly what we shall do here. We assign a *probability density* of 1 to each point in the interval $[0, 1]$. Then we determine the probability of a subinterval, $[1/3, 1/2]$ say, by calculating the integral $\mathbb{P}([1/3, 1/2]) = \int_{1/3}^{1/2} 1dx = 1/6$. More generally, the probability associated with an interval $[a, b]$ will just be its length $b - a$. The probability of a union of intervals will then be the sum of the lengths of those intervals, and so on.

This "continous" uniform distribution sometimes arises naturally from requirements of sym-

metry, just like its discrete counterpart. It can also arise as a limiting distribution. For instance, suppose that a hermit lives deep in a cave, away from any clocks or sources of natural light, and that each "day" he spends lasts for a random length of time between 23 and 25 hours. To start with, he will have some idea of what the time is, and be able to make statements such as, "I'm having lunch now, so it's probably light outside," but after a few weeks of this regime, he will no longer have any idea: any outside time will be just as likely as any other.

Now let us look at a rather more interesting density function, which depends on the choice of a positive constant λ. Consider the density function $f(x) = \lambda e^{-\lambda x}$, defined on the set of all non-negative real numbers. To work out the probability associated with an interval $[a, b]$, we now calculate

$$\int_a^b f(x)dx = \int_a^b \lambda e^{-\lambda x}dx = e^{-\lambda a} - e^{-\lambda b}.$$

The resulting probability distribution is called the *exponential distribution with parameter* λ. The exponential distribution is appropriate if we are modelling the time T of a spontaneous event, such as the time it takes for a radioactive nucleus to decay, or for the next spam email to arrive. The reason for this is based on the assumption of *memorylessness*: for example, if we know that the nucleus remains intact at time s, the probability that it will remain intact until a later time $s+t$ is the same as the original probability that it would remain intact to time t. Let $G(t)$ represent the probability that the nucleus remains intact up to time t. Then the probability that it remains intact up to time $s + t$ given that it has remained intact up to time s is $G(s+t)/G(s)$, so we are requiring that this equals $G(t)$. Equivalently, $G(s + t) = G(s)G(t)$. The only decreasing functions which have this property are EXPONENTIAL FUNCTIONS, that is, functions of the form $G(t) = e^{-\lambda t}$ for some positive λ. Since $1-G(t)$ represents the probability that the nucleus decays before time t, this should equal $\int_0^t f(x)dx$, from which it is easy to deduce that $f(x) = \lambda e^{-\lambda x}$.

We shall come to the third, and most important, distribution below.

4 Random variables, mean and variance

Given a probability space, an *event* is defined to be a (sufficiently nice) subset of that space. For example, if the probability space is the interval $[0, 1]$ with the uniform distribution, then the interval $[1/2, 1]$ is an event, which represents the result that a randomly chosen number between 0 and 1 is at least $1/2$. It is often useful to think not just about random events, but also about random *numbers* associated with a probability space. For example, let us look once again at a sequence of tosses of a biased coin that has probability p of coming up heads. The natural sample space associated with this experiment is the set Ω of all sequences ω of 0s and 1s. Earlier, we showed that the probability of obtaining k heads is $p_k = \binom{n}{k}p^k(1-p)^k$, and we described that as a distribution on the sample space $\{0, 1, 2, \dots, n\}$. However, it is in many ways more natural, and often far more convenient, to regard the original set Ω as the sample space and to define a function X from Ω to $\mathbb{R}$ to represent the number of heads: that is, $X(\omega)$ is the number of 1s in the sequence ω. We then write

$$\mathbb{P}(X = k) = p_k = \binom{n}{k}p^k(1-p)^k.$$

A function like this is called a *random variable*. If X is a random variable and it takes values in a set Y, then the *distribution* of X is the function defined on subsets of Y by $\mathbb{P}(A) = \mathbb{P}(X(\omega) \in A)$.

For many purposes, it is enough to know the distribution of a random variable. However, the notion of a random variable defined on a sample space captures our intuition of a random quantity, and it allows us to ask further questions. For example, if we were to ask for the probability that there were k heads, given that the first and last tosses had the same outcome, then the distribution of X would not provide the answer, whereas our richer model of regarding X as a function defined on sequences would do so. Furthermore, we can talk of *independent* random variables, $X_1, \dots, X_n$ say, meaning that the subset of Ω where $X_i(\omega) \in A_i$ for all i has probability given by the product $\mathbb{P}(X_1 \in A_1) \times \cdots \times \mathbb{P}(X_n \in A_n)$ for all possible sets of values A_i.

Associated with a random variable X are two important numbers that begin to characterize it,

called the *mean* or *expectation* $\mathbb{E}(X)$ and the *variance* $\mathrm{var}(X)$. Both these numbers are determined by the distribution of X. If X takes integer values, with distribution $\mathbb{P}(X = k) = p_k$, then

$$\mathbb{E}(X) = \sum_k k p_k, \quad \mathrm{var}(X) = \sum_k (k-\mu)^2 p_k,$$

where $\mu = \mathbb{E}(X)$. The mean tells us how big X is on average. The variance, or more precisely its square root, the *standard deviation* $\sigma = \sqrt{\mathrm{var}(X)}$, tells us how far away X lies, typically, from its mean. The useful formula $\mathrm{var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$ is an easy exercise.

As we discussed at the start of this article, it is known from experience that the "expected" number of heads in a sequence of n tosses of a fair coin is around $n/2$, in the sense that the proportion is usually close to $1/2$. It is not hard to work out that, if X models the number of heads in n tosses, that is, if X has binomial distribution of parameters n and $1/2$, then $\mathbb{E}(X) = n/2$. The variance of X is $n/4$, so the distribution is spread out on a scale of $\sigma = \sqrt{n}/2$. This allows us to see that X/n is close to $1/2$ with probability close to 1 for large n, in accordance with experience.

To understand the importance of the variance, consider the following situation. Suppose that 100 people take an exam and you are told that their average mark is 75%. This gives you some useful information, but by no means a complete picture of how the marks are distributed. For example, perhaps the exam consisted of four questions of which three were very easy and one almost impossible, so that all the marks were clustered around 75%. Or perhaps about 50 people got full marks and 50 got around half marks. To model this situation let the sample space Ω consists of the 100 people and let the probability distribution be the uniform distribution. Given a random person ω, let $X(\omega)$ be that person's mark. Then in the first situation, the variance will be small, since almost everybody's mark is close to the mean of 75%, whereas in the second it is close to $25^2 = 625$, since almost everybody's mark was about 25 from the mean. Thus, the variance helps us to understand the difference between the two situations.

A useful fact about variance is that if $X_1, X_2, \ldots, X_n$ are independent random variables, then $\mathrm{var}(X_1 + \cdots + X_n) = \mathrm{var}(X_1) + \cdots + \mathrm{var}(X_n)$. It follows that if all the X_i have the same distribution with mean μ and variance σ^2, then the variance of the *sample average* $\overline{X} = n^{-1}(X_1 + \cdots + X_n)$ is $n^{-2}(n\sigma^2) = \sigma^2/n$, which tends to zero as n tends to infinity. This observation can be used to prove that for any $\epsilon > 0$ the probability that $|\overline{X} - \mu| > \epsilon$ tends to zero as n tends to infinity. Thus, the sample average "converges in probability" to the mean μ.

This result is called the *weak law of large numbers*. The argument sketched above implicitly assumes that the random variables have finite variance, but this assumption turns out not to be necessary. There is also a *strong law of large numbers*, which states that, with probability 1, the sample average of the first n variables converges to μ as n tends to infinity. As its name suggests, the strong law is stronger than the weak law, in the sense that the weak law can be deduced from the strong law. Both laws give a rigorous backing for our initial notion that the probability of an event (such as a roll of a die coming up as a six) represents the average proportion of times that the event would occur over a large number of independent trials.

5 The normal distribution and the central limit theorem

As we have seen, for the binomial distribution with parameters p and n, the probability p_k is given by the formula $\binom{n}{k} p^k (1-p)^{n-k}$. If n is large and you plot the points (k, p_k) on a graph, then you will notice that they lie in a bell-shaped curve that has a sharp peak around the mean np. The width of the tall part of the curve has order of magnitude $\sqrt{np(1-p)}$, the standard deviation of the distribution. Let us assume for simplicity that np is an integer, and define a new probability distribution q_k by $q_k = p_{k+np}$. The points (k, q_k) peak at $k = 0$. If you now rescale the graph, compressing horizontally by a factor of $\sqrt{np(1-p)}$ and expanding vertically by the same factor, then the points will all lie close to the graph of

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

This is the density function of a famous distribution known as the *standard normal distribution* on $\mathbb{R}$.

To put this differently, if you toss a biased coin a large number of times, then the number of heads, minus the mean and divided by the standard deviation, is close to a standard normal random variable.

The function $\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ occurs in a huge variety of mathematical contexts, from probability theory to FOURIER ANALYSIS to quantum mechanics. Why should this be? The answer, as it is for many such questions, is that there are properties that this function has that are shared by no other function.

One such property is *rotational invariance*. Suppose once again that we are throwing a dart at a dartboard and aiming for the bull's eye. We could model this as the result of adding two independent normal distributions at right angles to each other: one for the x-coordinate and one for the y-coordinate (say each having mean 0 and variance 1). This would make the two-dimensional "density function" be $\frac{1}{2\pi}e^{-x^2/2}e^{-y^2/2}$, which can conveniently be written as $\frac{1}{2\pi}e^{-r^2/2}$, where r denotes the length of (x, y). In other words, the density function depends only on the distance from the origin. (This is why it is called "rotationally invariant.") This very appealing property holds in more dimensions as well. And it turns out to be quite easy to check that $\frac{1}{2\pi}e^{-r^2/2}$ is the *only* such function: more precisely, it is the only rotation-invariant density function (of variance 1, say) that can be decomposed into separate independent random variables of x and y. Thus, the normal distribution has a very special symmetry property.

Properties like this go some way towards explaining the ubiquity of the normal distribution in mathematics. However, the normal distribution has an even more remarkable property, which makes it ubiquitous in *nature*. The *central limit theorem* states that, for any sequence of independent and identically distributed random variables $X_1, X_2, \dots$ (with finite mean μ and non-zero finite variance σ^2), we have

$$\lim_{n\to\infty} \mathbb{P}(X_1+\cdots+X_n \leq n\mu+\sqrt{n}\sigma x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}}e^{-y^2/2}dy$$

for every real number x. The expected value of $X_1 + \cdots + X_n$ is $n\mu$ and its standard deviation is $\sqrt{n}\sigma$, so another way of thinking about this is to let $Y_n = (X_1 + \cdots + X_n - n\mu)/\sqrt{n}\sigma$. This rescales $X_1 + \cdots + X_n$ to have mean 0 and variance 1, and the probability becomes the probability that $Y_n \leq x$. Thus, *whatever* distribution we start with, the limiting distribution (after appropriate rescaling) is normal. Many natural processes can realistically be modelled as accumulations of small independent random effects, and this is why many distributions one observes, such as the distribution of heights of adults in a given town, have a familiar bell-shaped curve.

A very useful application of the central limit theorem is to simplify what look like impossibly complicated calculations. For example, when the paramenter n is large, the calculation of binomial probabilities becomes prohibitively complicated, but we may instead write a binomial random variable X, of parameters n and say $1/2$, in the form $X = Y_1 + \cdots + Y_n$, with $Y_1, \dots, Y_n$ independent Bernoulli of parameter $1/2$. Then, by the central limit theorem,

$$\lim_{n\to\infty} \mathbb{P}(X \leq n/2 + \sqrt{n}x/2) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}}e^{-y^2/2}dy.$$