Probabilities and Random Variables

This is an elementary overview of the basic concepts of probability theory.

1 The Probability Space

The purpose of probability theory is to model random experiments so that we can draw inferences about them. The fundamental mathematical object is a triple $(\Omega, \mathcal{F}, P)$ called the *probability space.* A probability space is needed for each experiment or collection of experiments that we wish to describe mathematically. The ingredients of a probability space are a *sample space* Ω, a collection $\mathcal{F}$ of *events*, and a *probability measure* P. Let us examine each of these in turn.

1.1 The sample space Ω

This is the set of all the possible outcomes of the experiment. Elements of Ω are called *sample points* and typically denoted by ω. These examples should clarify its meaning:

Example 1. If the experiment is a roll of a six-sided die, then the natural sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$. Each sample point is a natural number between 1 and 6.

Example 2. Suppose the experiment consists of tossing a coin three times. Let us write 0 for heads and 1 for tails. The sample space must contain all the possible outcomes of the 3 successive tosses, in other words, all triples of 0's and 1's:

$$\begin{aligned}\Omega &= \{0,1\}^3 \\ &= \{0,1\} \times \{0,1\} \times \{0,1\} \\ &= \{(x_1, x_2, x_3) : x_i \in \{0,1\} \text{ for } i = 1,2,3\} \\ &= \{(0,0,0), (0,0,1), (0,1,0), (0,1,1), (1,0,0), (1,0,1), (1,1,0), (1,1,1)\}.\end{aligned}$$

The four formulas are examples of different ways of writing down the set Ω. A sample point $\omega = (0, 1, 0)$ means that the first and third tosses come out heads (0) and the second toss comes out tails (1).

Example 3. Suppose the experiment consists of tossing a coin infinitely many times. Even though such an experiment cannot be physically arranged, it is of central importance to the theory to be able to handle idealized situations of this kind. A sample point ω is now an infinite sequence

$$\omega = (x_1, x_2, x_3, \ldots, x_i, \ldots) = (x_i)_{i=1}^{\infty}$$

whose terms x_i are each 0 or 1. The interpretation is that $x_i = 0$ (1) if the ith toss came out heads (tails). In this model the infinite sequence of coin tosses is regarded as a single experiment, although we may choose to observe the individual tosses one at a time. (If you need a mental picture, think of the goddess of chance simultaneously tossing all infinitely many coins.) The sample space Ω is the space of all infinite sequences of 0's and 1's:

$$\begin{aligned}\Omega &= \{\omega = (x_i)_{i=1}^{\infty} : \text{ each } x_i \text{ is either 0 or 1 }\} \\ &= \{0,1\}^{\mathbb{N}}.\end{aligned}$$

In the last formula $\mathbb{N} = \{1, 2, 3, \ldots\}$ stands for the set of natural numbers, and the notation $\{0,1\}^{\mathbb{N}}$ stands for the infinite product set

$$\{0,1\}^{\mathbb{N}} = \{0,1\} \times \{0,1\} \times \{0,1\} \times \cdots \times \{0,1\} \times \cdots .$$

Example 4. If our experiment consists in observing the number of customers that arrive at a service desk during a fixed time period, the sample space should be the set of nonnegative integers: $\Omega = \mathbb{Z}_+ = \{0, 1, 2, 3, \ldots\}$.

Example 5. If the experiment consists in observing the lifetime of a light bulb, then a natural sample space would be the set of nonnegative real numbers: $\Omega = \mathbb{R}_+ = [0, \infty)$.

These sample spaces are mathematically very different. Examples 1 and 2 have *finite* sample spaces, while the rest are *infinite*. The sample space in Example 4 is *countably infinite*, which means that its elements can be arranged in a sequence. Finite and countably infinite spaces are also called *discrete*. Probability theory on discrete sample spaces requires no advanced mathematics beyond calculus and linear algebra, and that is the main focus of this course. A precise treatment of probability models on *uncountable* spaces requires measure theory.

1.2 The collection of events, $\mathcal{F}$

Events are simply subsets of the sample space. They can often be described both in words and in set-theoretic notation. Events are typically denoted by upper case letters: A, B, C, etc. Here are some possible events on the sample spaces described above:

Example 1.

$$A = \{\text{the outcome of the die is even}\} = \{2, 4, 6\}.$$

Example 2.

$$A = \{\text{exactly two tosses come out tails}\} = \{(0,1,1), (1,0,1), (1,1,0)\}.$$

Example 3.

$$\begin{aligned} A &= \{\text{the second toss comes out tails and the fifth heads}\} \\ &= \{\omega = (x_i)_{i=1}^{\infty} \in \{0,1\}^{\mathbb{N}} : \ x_2 = 1 \text{ and } x_5 = 0 \ \}. \end{aligned}$$

Example 4.

$$A = \{\text{at least 6 customers arrived}\} = \{6, 7, 8, 9, \ldots\}.$$

Example 5.

$$A = \{\text{the bulb lasted less than 4 hours}\} = [0, 4).$$

In discrete sample spaces the class $\mathcal{F}$ of events contains *all subsets of the sample space.* But in more complicated sample spaces this cannot be the case. In the general theory $\mathcal{F}$ is a *σ-algebra* in the sample space. This means that $\mathcal{F}$ satisfies the following axioms:

(i) $\Omega \in \mathcal{F}$ and $\emptyset \in \mathcal{F}$ (The sample space Ω and the empty set $\emptyset$ are events.)

(ii) If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$. (A^c is the complement of A, this is the set of points of Ω that do not belong in A.)

(iii) If $A_1, A_2, A_3, \ldots \in \mathcal{F}$ then also $\displaystyle\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$. (In words: the union of a sequence of events is also an event.)

1.3 The probability measure P

This is a function on events that gives the probability $P(A)$ of each event A in $\mathcal{F}$. It must satisfy these axioms:

(i) $0 \le P(A) \le 1$ for all $A \in \mathcal{F}$.

(ii) $P(\emptyset) = 0$ and $P(\Omega) = 1$.

(iii) If $A_1, A_2, A_3, \ldots$ are pairwise disjoint events, meaning that $A_i \cap A_j = \emptyset$ whenever $i \neq j$, then

$$P\Bigl(\bigcup_{i=1}^{\infty} A_i\Bigr) = \sum_{i=1}^{\infty} P(A_i).$$

The necessity to restrict $\mathcal{F}$ in uncountable sample spaces arises from the fact that we cannot always consistently define probabilities for all subsets of uncountable sample spaces. The way around this difficulty is to admit as events only members of certain judiciously defined σ-algebras. These complications cannot be appreciated without some study of measure theory. The good news is that in practice one never encounters these bad events that the theory cannot handle. It is a hard exercise in real analysis to construct a 'nonmeasurable set' on $\mathbb{R}$ which lies outside the natural σ-algebra.

In discrete sample spaces these difficulties can be completely avoided. Suppose our sample space Ω is either finite or countably infinite. To define a probability measure on Ω we only need an assigment $P(\omega)$ of probabilities for all sample points ω that together satisfy

$$0 \le P(\omega) \le 1 \text{ for all } \omega \in \Omega$$

and

$$\sum_{\omega \in \Omega} P(\omega) = 1.$$

Then for any subset A of Ω we can define a probability by

$$P(A) = \sum_{\omega \in A} P(\omega). \tag{1}$$

This concrete construction shows that there is no problem in giving each subset $A \subset \Omega$ a probability, and it is easy to check that axioms (i)–(iii) for P above are satisfied.

This approach does not work for uncountable spaces such as the real line. The interesting probability measures there typically give individual sample points zero probability, and the summation in (1) has to be replaced by integration.

Here are examples of natural probability measures on the previously defined sample spaces:

Example 1. If we assume that the die is fair, then it is reasonable to regard each outcome as equally likely, and the appropriate probability measure is $P(\omega) = 1/6$ for each $\omega \in \Omega$.

Example 2. If the coin is fair, then each outcome should be equally likely: $P(\omega) = 1/8$ for each $\omega \in \Omega$. If the coin is not fair but gives tails with probability p and heads with probability $1 - p$ for some fixed number $p \in (0, 1)$, then the appropriate P is defined for sample points $\omega = (x_1, x_2, x_3)$ by

$$P(x_1, x_2, x_3) = p^{x_1+x_2+x_3}(1-p)^{3-(x_1+x_2+x_3)}.$$

The uniform measure $P(\omega) = 1/8$ is recovered in the case $p = 1/2$.

Example 3. Defining a probability measure on the space of infinite sequences requires some measure-theoretic machinery. We will not concern ourselves with these technical issues. We simply take for granted that it is possible to construct the probability measures on Ω that we need. For example, there is a probability measure P on Ω such that, for any choice of numbers $a_1, a_2, \ldots, a_n$ from $\{0, 1\}$,

$$P\{\omega \in \Omega : \omega_1 = a_1, \ldots, \omega_n = a_n\} = 2^{-n}. \tag{2}$$

The number 2^{-n} corresponds to a fair coin ($p = 1/2$ in Example 2), and the construction can be performed for any p just as well. The result from measure theory that makes all this work is called *Kolmogorov's extension theorem.*

Example 4. A natural choice of probability measure here could be a Poisson distribution:

$$P(\omega) = \frac{e^{-\lambda}\lambda^{\omega}}{\omega!} \qquad \text{for } \omega = 0, 1, 2, 3, \ldots \tag{3}$$

The number $\lambda > 0$ is a parameter of the model that in an actual modeling task would be chosen to fit the observed data.

Example 5. A possible choice here would be an exponential distribution with parameter $\lambda > 0$. This can be defined by saying that the probability of an interval $[a, b]$ with $0 \le a < b < \infty$ is given by

$$P[a, b] = \int_a^b \lambda e^{-\lambda x} dx.$$

Here $f(x) = \lambda e^{-\lambda x}$ is the *density function* of the exponential distribution with parameter λ. Equivalently, the same probability measure can be defined by giving the *(cumulative) distribution function*

$$F(x) = P[0, x] = 1 - e^{-\lambda x}, \quad x \ge 0.$$

2 Random Variables and Their Expectations

From now on the discussion always takes place in the context of some fixed probability space $(\Omega, \mathcal{F}, P)$. You may assume that Ω is discrete, that $\mathcal{F}$ is the collection of all subsets of Ω, and that P is defined by (1) from some given numbers $\{P(\omega) : \omega \in \Omega\}$ that give the probabilities of individual sample points.

A *random variable* is a function defined on the sample space Ω. While functions in calculus are typically denoted by letters such as f or g, random variables are often denoted by capital letters such as X, Y and Z. Thus a random variable X associates a number $X(\omega)$ to each sample point ω.

For each event A there is an indicator random variable I_A defined by

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

Other common notations for an indicator random variable are 1_A and χ_A.

If $X_1, X_2, \dots, X_n$ are random variables defined on some common probability space, then $X = (X_1, X_2, \dots, X_n)$ defines an $\mathbb{R}^n$-valued random variable, also called a *random vector*.

Most interesting events are of the type $\{\omega \in \Omega : X(\omega) \in B\}$, where X is a random variable and B is a subset of the space into which X maps (so typically the real line $\mathbb{R}$ or, in the case of a random vector, a Euclidean space $\mathbb{R}^n$). The above notation is usually abbreviated as $\{X \in B\}$, and it reads "the event that the value of X lies in the set B."

Examples:

Example 2. Let the random variable Y be the outcome of the third toss. If we denote the elements of Ω by $\omega = (x_1, x_2, x_3)$, then Y is defined by $Y(\omega) = x_3$. Suppose you receive \$5 each time the coin comes up heads. Let Z be the amount you won. Then you could express Z as a function on Ω by writing $Z(\omega) = 5(3 - (x_1 + x_2 + x_3))$. The event $\{Z = 10\}$ is the subset of sample points ω that satisfy the condition $Z(\omega) = 10$. So

$$\{Z = 10\} = \{\omega \in \Omega : Z(\omega) = 10\} = \{(0,0,1), (0,1,0), (1,0,0)\}.$$

The probability measure P on the sample space gives the probabilities of the values of a random variable. For example, in the case of a fair coin,

$$P\{Z = 10\} = P\{(0,0,1), (0,1,0), (1,0,0)\} = 3/8.$$

Example 3. In the sequence setting it is customary to use these so-called coordinate random variables X_i: For a sample point $\omega = (x_i)_{i=1}^{\infty}$, define $X_i(\omega) = x_i$. If our probability measure P is the one given by formula (2), then for any choice of numbers $a_1, a_2, \dots, a_n$ from $\{0, 1\}$,

$$P\{X_1 = a_1, \dots, X_n = a_n\} = 2^{-n}. \tag{4}$$

If S_n is the number of tails in the first n tosses, then

$$S_n = \sum_{i=1}^{n} X_i. \tag{5}$$

Here we are defining a new random variable S_n in terms of the old ones X_1, X_2, X_3, … The sample point has been left out of formula (5) as is typical to simplify notation. The precise meaning of equation (5) is that for each $\omega \in \Omega$, the number $S_n(\omega)$ is defined in terms of the numbers $X_1(\omega)$, $X_2(\omega)$, $X_3(\omega)$, … by

$$S_n(\omega) = \sum_{i=1}^{n} X_i(\omega). \tag{6}$$

Again, probabilities of the values of random variables come from the probability measure on the sample space. From (2) or (4) it follows that

$$\begin{aligned} P\{S_5 = 3\} &= \text{(the number of ways of getting 3 tails in 5 tosses)} \cdot 2^{-5} \\ &= \binom{5}{3} 2^{-5} = \frac{5}{16}. \end{aligned}$$

Example 4. Let the random variable X be the number of customers. Then, as a function on Ω, X is somewhat trivial, namely the identity function: $X(\omega) = \omega$. Formula (3) from above can be expressed as

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!} \qquad \text{for } k = 0, 1, 2, 3, \ldots \tag{7}$$

Now we say that the random variable X is Poisson distributed with parameter λ.

Examples 3 and 4 are instructive in the sense that we prefer to express things in terms of random variables. That is, we prefer formulas (4) and (7) over formulas (2) and (3). The probability space recedes to the background and sometimes vanishes entirely from the discussion. In example 4 we would simply say "let X be a random variable with Poisson(λ) distribution" by which mean that (6) holds. No probability space needs to be specified because it is understood that the simple probability space of Example 4 can be imagined in the background.

No matter what the original probability space $(\Omega, \mathcal{F}, P)$ is, we can often translate all the relevant probabilistic information to a more familiar space such as $\mathbb{R}$ or $\mathbb{Z}_+$ by using the distribution of a random variable. Suppose first that X is a $\mathbb{Z}_+$-valued random variable. (This means that X is a function from Ω into $\mathbb{Z}_+$.) The *distribution* (or *probability distribution*) of X is a sequence $(p_k)_{k=0}^{\infty}$ of numbers defined by

$$p_k = P(X = k), \qquad k = 0, 1, 2, 3, \ldots \tag{8}$$

Thus the distribution of X is actually a probability measure on $\mathbb{Z}_+$. All probabilistic statements about X can be expressed in terms of its distribution, so we can forget the original probability space. Sometimes we need to allow for the possibility that X takes on the value ∞. Then an additional number

$$p_\infty = P(X = \infty) = 1 - \sum_{k=0}^{\infty} p_k$$

needs to be added to the distribution.

If we want to talk about more than one random variable simultaneously, we need joint distributions. Suppose $X_1, X_2, \ldots, X_n$ are $\mathbb{Z}_+$-valued random variables defined on a common probability space. Their *joint distribution* is a collection

$$(p_{k_1,\ldots,k_n} : k_1, \ldots, k_n \in \mathbb{Z}_+)$$

of numbers indexed by the n-tuples $(k_1, \ldots, k_n)$ of nonnegative integers, and defined by

$$p_{k_1,\ldots,k_n} = P(X_1 = k_1, \ldots, X_n = k_n).$$

An example: formula (4) above specifies the joint distribution of the first n coin tosses.

Similar formulas are valid when the range of X is some other countable subset of $\mathbb{R}$ instead of $\mathbb{Z}_+$. But when the range of X is not countable, a sequence of numbers such as (8) cannot specify its distribution. Then we need the *(cumulative) distribution function (c.d.f.)* of X, denoted by F and defined for all real numbers x by

$$F(x) = P(X \leq x).$$

If F is differentiable, then $f(x) = F'(x)$ is the *probability density* of X.

The *expectation* of a random variable X is a number defined by

$$E[X] = \sum_x xP(X = x) \tag{9}$$

if X has only countably many possible values x, and by

$$E[X] = \int_{-\infty}^{+\infty} xf(x)dx$$

if X has density f, provided the sum (integral) exists. Both formulas can be subsumed under the formula

$$E[X] = \int_{-\infty}^{+\infty} x\, dF(x),$$

where F is the c.d.f. of X, and the integral is interpreted as a Riemann-Stieltjes integral. (The Riemann-Stieltjes integral is a topic of a course on advanced calculus or beginning real analysis. Do not worry even if you are not familiar with it.)

It follows from formula (9) that if X is $\mathbb{Z}_+$-valued, then

$$E[X] = \sum_{k=0}^{\infty} kp_k,$$

and furthermore, if f is a function defined on $\mathbb{Z}_+$, then the composition $f(X)$ is a new random variable and it has expectation

$$E[f(X)] = \sum_{k=0}^{\infty} f(k)p_k.$$

Note that $E[X] = \infty$ if $P(X = \infty) > 0$.

The expectation of an indicator random variable is the probability of the event in question, as an easy calculation shows:

$$E[I_A] = 1 \cdot P(I_A = 1) + 0 \cdot P(I_A = 0) = 1 \cdot P(A) + 0 \cdot P(A^c) = P(A).$$

Example 2. Let us calculate the expectation of the reward Z, assuming the coin is fair:

$$\begin{aligned}
E[Z] &= 0 \cdot P(Z = 0) + 5 \cdot P(Z = 5) + 10 \cdot P(Z = 10) + 15 \cdot P(Z = 15) \\
&= 5 \cdot \binom{3}{1} \cdot \frac{1}{8} + 10 \cdot \binom{3}{2} \cdot \frac{1}{8} + 15 \cdot \binom{3}{3} \cdot \frac{1}{8} \\
&= \frac{5 \cdot 3}{8} + \frac{10 \cdot 3}{8} + \frac{15 \cdot 1}{8} \\
&= \frac{15}{2}.
\end{aligned}$$

An important property of the expectation is its *additivity*: Suppose $X_1, \dots, X_n$ are random variables defined on a common probability space and $c_1, \dots, c_n$ are numbers. Then $c_1X_1 + \cdots + c_nX_n$ is also a random variable and

$$E[c_1X_1 + \cdots + c_nX_n] = c_1E[X_1] + \cdots + c_nE[X_n].$$

3 Independence and Conditioning

Suppose $X_1, \dots, X_n$ are discrete random variables (they have countable ranges) defined on a common probability space. They are *mutually independent* if the following holds for all possible values $y_1, \dots, y_n$ of their range:

$$P(X_1 = y_1, X_2 = y_2, \dots, X_n = y_n) = P(X_1 = y_1) \cdot P(X_2 = y_2) \cdots P(X_n = y_n). \tag{10}$$

A collection of events $A_1, \dots, A_n$ are mutually independent if their indicator random variables are mutually independent.

Given an event B such that $P(B) > 0$, we can define a new probability measure $P(\cdot\,|B)$ on Ω by *conditioning on* B: The conditional probability of an event A, given B, is defined by

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

You can think of conditioning as first restricting the sample space to the set B, and then renormalizing the probability measure by dividing by $P(B)$ so that the total (new) space again has probability 1.

Check that the following holds: Two events A and B are independent if and only if $P(A|B) = P(A)$ and $P(B|A) = P(B)$. The equality $P(A|B) = P(A)$ means that the event B gives no information about the event A, in the sense that the probability of A's occurrence is not influenced by whether or not B occurred. This then is the intuitive meaning of statistical independence: If X and Y are independent random variables, knowledge of the value of X should give us no information about the value of Y.

Independence splits the expectation of a product of random variables into a product of expectations: Suppose $X_1, \dots, X_n$ are mutually independent random variables whose expectations are finite. Then the product $X_1 \cdot X_2 \cdot X_3 \cdots X_n$ is a random variable, and

$$E[X_1 \cdot X_2 \cdot X_3 \cdots X_n] = E[X_1] \cdot E[X_2] \cdot E[X_3] \cdots E[X_n].$$