10 Introduction to stochastic processes and probability

Models and methods that have been considered so far in the course have been deterministic — a single input produces the same answer every time while solutions to problems are aimed at finding one correct answer or a close approximations to it. These deterministic models, methods and approximations can be very accurate, particularly in engineering and physical applications.

Many systems, however, are inherently random. Apparently identical inputs may produce radically different outputs and no two realisations of the system are exactly the same. Whether that randomness is the result of our imprecise measurements (were the inputs exactly the same?) or there is a fundamental randomness built into the system (quantum uncertainty?), we should attempt to model and quantify this uncertainty. We regularly refer to non-deterministic systems as *stochastic* rather than random to avoid the common usage of random (where it is often used to mean uniformly random where every outcome is equally likely).

The framework we use to make these models is probability theory and, ideally, we use statistical inference to find the relationship between our models and the system we are studying. Simulation is one of the tools we use to understand how our models behave and is often used where exact statistical inference is prohibitively difficult. Both statistical inference and simulation rely heavily on computational power and algorithms. Model building is typically more of an art than a science and is done by hand (or mind).

In this section, we look at some of the basic terminology of probability theory, introduce the fundamental ideas behind statistical inference and see how we can simulate stochastic processes *in silico*.

11 Primer on Probability

The basic challenge of probability theory and applied probability is to understand and describe the laws according to which events occur.

A event can be pretty much anything. Commonly used examples in probability are rolling dice, picking balls out of an urn or tossing a coin — these are commonly used because they are simple, easy to understand and aid our intuition. But all sorts of events can be thought of as random: the amount of rain falling in an area in a given period, the number of mutations that occur when a cell splits, the age of the person currently reading this sentence. Indeed, if we consider randomness to a a property of our state of knowledge of an event, any event can be considered random.

Formally, we define a *probability* to be a number between 0 and 1 assigned to a set of outcomes of a random process called an *event*. This number is typically interpreted as the chance of the event occurring or as the degree of plausibility we place on the event occurring.

The set of all outcomes that the random process can take is known as the *state space*

and is often denoted Ω.

An *event*, A, is a subset of the state space: $A \subseteq \Omega$. When Ω is finite or countable, probability can be viewed as a function from the set of all subsets of Ω, written $\mathcal{A}$, to the interval $[0, 1]$, that is $P : \mathcal{A} \to [0, 1]$. $\mathcal{A}$, the set of all subsets of Ω, is called the *power set* of Ω. That is, for some event or collection of events, we view the function P as giving the probability of that event occurring.

This interpretation is not mathematically correct when Ω is uncountable (for example, when our random process can take any value in continuous interval) but it will be sufficient to guide our intuition here.

Example: Tossing a coin 2 times and recording the result of each toss. The random process is tossing the coin twice. The state space is the set of all possible outcomes: $\Omega = \{HH, HT, TH, TT\}$.

An example of an event is that we throw a tails first: in this case $A = \{TH, TT\}$.

There are $2^4 = 16$ possible events that we could consider here as that is the size of the power-set (the set of all possible subsets) of Ω. For completeness, we write down all possible events: $\mathcal{A} = \{\emptyset, \{HH\}, \{HT\}, \{HT\}, \{TT\}, \{HH, HT\}, \{HH, TH\}, \{HH, TT\}, \{HT, TH\}, \{HT, TT\}, \{TH, TT\}, \{HH, HT, TH\}, \{HH, HT, TT\}, \{HH, TH, TT\}, \{HT, TH, TT\}, \Omega\}$. □

Note that sometimes we distinguish between simple and compound events. In the above example, the simple events are $\{HH\}, \{HT\}, \{HT\}, \{TT\}$ while compound events are combinations of simple events.

Example: Length of time waiting for bus, measured from arrival at bus stop until bus arrives. Supposing the buses come every 15 mins. Then $\Omega = [0, 15]$ (that is, any time in the interval between 0 and 15 minutes). An example of an event is $A = [0, 1]$ being the event that the wait for the bus is at most 1 minute. □

Let A and B be events. Then the event C that A and B occur is given by $C = A \cap B$ while the event F that A or B occurs is given by $D = A \cup B$. The event A does not occur is given by $A^c = \bar{A} = \Omega - A = \Omega \backslash A = \{\omega \in \Omega : \omega \notin A\}$.

Example: Suppose we roll a fair die and record the value. Then $\Omega = \{1, 2, 3, 4, 5, 6\}$. Let A be the event that the roll is even, B be the event that we roll a 3 or a 6. Then $A = \{2, 4, 6\}$ and $B = \{3, 6\}$. The event that A and B occur is $A \cap B = \{6\}$ while the event that A or B happens is $A \cup B = \{2, 3, 4, 6\}$. The event that A does not occur is $A^c = \{1, 3, 5\}$ (A does not occur when the roll is odd). □

11.1 Axioms of probability

Any probability function must satisfy the 3 axioms (rules) of probability. The axioms are:

1. $P(\Omega) = 1$. That is, the total probability is 1.
2. $0 \leq P(A) \leq 1$ for any $A \subseteq \Omega$. The probability of any event is non-negative and

less than or equal to 1.

3. If $A_1, A_2, \ldots$ are mutually disjoint events (i.e., $A_i \cap A_j = \emptyset$ if $i \neq j$) then

$$P\left(\bigcup_i A_i\right) = \sum_i P(A_i).$$

From the above axioms, all the useful rules of probability can be derived. For example, $P(A^c) = 1 - P(A)$ since $1 = P(\Omega)$ (axiom 1) $= P(A \cup A^c)$ (definition of A^c) $= P(A) + P(A^c)$ (axiom 3 as A and A^c are disjoint).

11.2 Conditional probability and independence

For events A and B, if we know that B occurred, what can we say about the probability of A given that knowledge? This is captured by the concept of the *conditional probability of A given B* is written $P(A|B)$ and defined by

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

This is only defined where $P(B) > 0$.

Example: Suppose we roll a fair die and record the value. Let A be the event that 2 is rolled and B be the event that the roll is an even number. What is the probability that the roll is a two given that we know it is even? This is just $P(A|B)$. We calculate it as follows. $P(B) = 1/2$ and $P(A \cap B) = P(A) = 1/6$ since $A \cap B = A$. Thus

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/6}{1/2} = \frac{1}{3}.$$

□

We say that events A and B are *independent* when $P(A \cap B) = P(A)P(B)$. From the definition of conditional probability, it is clear that if A and B are independent, then $P(A|B) = P(A)$ and $P(B|A) = P(A)$.

Note that we usually write $P(A, B)$ instead of $P(A \cap B)$. More generally, we write $P(A_1, \ldots, A_k)$ for $P(\bigcap_{i=1}^k A_i)$.

Rearranging the definition of conditional probability, we see that $P(A, B) = P(A|B)P(B) = P(B|A)P(B)$. Repeated applications of this result gives

$$P(A_1, \ldots, A_k) = P(A_1|A_2, \ldots, A_k)P(A_2|A_3, \ldots, A_k) \ldots P(A_{k-1}|A_k)P(A_k).$$

11.3 Bayes' Theorem

From the definition of conditional probability, we can prove the following result, known as Bayes' theorem.

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

This simple result is important because it tells us how the forward probability $P(A|B)$ is related to the backward probability $P(B|A)$. We'll see that this relationship is crucial to statistical inference.

11.4 Random variables

A *random variable* (r.v.) X is a variable whose value results from the measurement of a random process. That is, a random variable is a measurement of some random event. We use capital letters to denote random variables while lower-case letters to denote particular observations or *realisations* of the random variable. So $X = x$ is the event that the random variable X takes the particular value x.

A **discrete** random variable takes a finite or countably infinite number of values, while a **continuous** random variable can take an uncountable number of possible values.

Random variables are most commonly real valued (that is, their value is a real number) but they can take any value. For example, we could consider random sequences, random graphs or random trees. For now, lets stick with real valued random variables. Formally, a real-valued random variable is a map from events to the real numbers: $X : \Omega \to \mathbb{R}$.

For discrete random variables, the **probability distribution function** (pdf) or **probability mass function** (pmf) is a function (rule, table) that assigns probabilities to each possible value of X.

$P(X = x) = p(x)$ is the probability that $X = x$. Sometimes write $P(X = x) = p_x$ or $P(X = x) = f(x)$.

As usual, we have $0 \leq P(X = x) \leq 1$ and $\sum_x P(X = x) = 1$.

For continuous random variables, the probability that a random variable takes any one exact value is zero, that is $P(X = x) = 0$, so we consider instead the **probability density function**, $p_X(x)$ (also written $f_X(x)$ or $f(x)$) from which we can calculate the probability that X lies in the interval $[a, b]$:

$$Pr(a \leq X \leq b) = \int_a^b p_X(x)dx.$$

$p_X(x)$ is real-valued, non-negative and normalised, i.e.,

$$\int_{-\infty}^{\infty} p_x(x)dx = 1.$$

Note that the integral $\int_a^b p_X(x)dx$ gives the area under the curve $p_X(x)$ between $x = a$ and $x = b$.

The **cumulative distribution function** (cdf), or simply the *distribution function*, is defined, for both discrete and continuous random variables, by $F(x) = P(X \leq x)$. This is a function that is monotonically increasing from 0 to 1. For continuous random variables, the cdf is continuous, while for discrete random variables, it is a step function with dis-continuities.

These ideas immediately extend to multiple random variables, so that the **joint probability density function** of n random valuables $X_1, \ldots, X_n$ takes n arguments, $p_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$ that is real-valued, non-negative and normalised. The probability that the point $(X_1, \ldots, X_n)$ lies in some region is just the multiple integral over that region.

Given a joint probability density function, we obtain the probability density for a subset of the variables by integrating over the ones not in the subset. For example, given $p_{XY}(x, y)$, we have

$$p_X(x) = \int_{-\infty}^{\infty} p_{XY}(x, y)dy.$$

This process is known as **marginalization**. The process is the same for a discrete variable, if we replace the integral with a sum:

$$P(x) = \sum_{y \in \mathcal{Y}} P(x, y)$$

Two random variables X and Y are **independent** when

$$p_{XY}(x, y) = p_X(x)p_Y(y).$$

Equivalently, X and Y are independent when

$$p_{Y|X}(y|x) = p_Y(y).$$

The **expected value** of a random variable X is called the mean and is given by

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

For discrete random variables, this is written

$$E[X] = \sum_{x \in \mathcal{X}} xp_x.$$

The symbol μ is often used for the mean.

The **variance** of a random variable is $\mathrm{Var}(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2$. The variance is a measure of the spread of a random variable about its mean.

The **expected value of a function f of X** is

$$E[f(X)] = \int_{-\infty}^{\infty} f(x)p_X(x)dx.$$