Lecture 1

1 Probability axioms

Example 1: Roll a die. Suppose the outcomes $1, \ldots, 6$, have probabilities $1/4, 1/4, 1/8, 1/8, 1/8, 1/8$ respectively. What is the probability of (a) an even number, (b) a prime number?

Example 2: Same experiment but we are now given the following information:

$$P(\{1,2\}) = P(\{3,4\}) = P(\{5,6\}) = 1/3, \quad P(\{1,2,3\}) = 1/2.$$

What is (a) $P(\{3\})$, (b) P({ 4 }), (c) P({ 6 })?

Example 3: Same experiment but we are now given the following information:

$$P(\{1,2\}) = P(\{3,4\}) = P(\{5,6\}) = 1/2, \quad P(\{1\}) = 1/4.$$

What is $P(\{2\})$?

Formalise this intuition.

The sample space Ω is an arbitrary set, thought of as the set ofpossible outcomes. Events are subsets of the sample space, and probabilities are numbers assigned to events. The collection of events, $\mathcal{F}$, to which we assign probabilities has a certain structure:

1. $\Omega \in \mathcal{F}$
2. If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$.
3. If $A_n \in \mathcal{F}$ for $n = 1, 2, 3, ...$, then $\cup_{n=1}^{\infty} A_n \in \mathcal{F}$.

A collection of subsets of Ω having these properties is called a σ-algebra.

Probability axioms: A function $P : \mathcal{F} \to \mathbb{R}$ is called a probability if

1. $0 \leq P(A) \leq 1$ for all $A \in \mathcal{F}$, and $P(\Omega) = 1$.
2. If $A_1, A_2, \ldots$ are disjoint sets, then $P(\cup_{n=1}^{\infty} A_n = \sum_{n=1}^{\infty} P(A_n)$.

Example 4: A dart is thrown at a circular dartboard of radius 1m and is equally likely to land anywhere on it. (What does this mean?) How likely is it to land in (a) the top half, (b) the bull's eye, which is a central circle of radius 1cm, (c) the point with co-ordinates $(+0.2, -0.3)$, (d) the horizontal diameter?

The point of this example is to illustrate why it is not possible to extend the axioms to uncountable sums.

2 Conditional probability and independence

Example: Let's go back to the example of rolling a die, where the outcomes $1, \ldots, 6$, have probabilities $1/4, 1/4, 1/8, 1/8, 1/8, 1/8$ respectively. Given that an even number was rolled, how likely are the events (a) $\{2\}$, (b) $\{3, 4, 5\}$?

Definitions: For events A and B, if $P(B) > 0$, then $P(A|B) := P(A \cap B)/P(B)$. Note: $P(\cdot|B)$ is a probability.

We say that events A and B are independent of each other if $P(A \cap B) = P(A)P(B)$. If $P(B) > 0$, this is the same as saying that $P(A|B) = P(A)$.

Events $A_1, \ldots, A_n$ are mutually independent if

$$P(\cap_{i=1}^{n} A_i) = \prod_{i=1}^{n} P(A_i). \tag{1}$$

An infinite sequence of events is said to be mutually independent if every finite subcollection of them is mutually independent. (Why would it be a bad idea to define it analogous to (1)?)

Total probability formula and Bayes' formula: Let $A_1, A_2 \ldots, A_n$ be a partition of Ω, i.e., the sets are mutually disjoint and their union is Ω.

(I mean measurable partition, but I'll omit the qualifier henceforth on the grounds that we will only consider measurable sets.) Then, for any event B,

$$P(B) = \sum_{i=1}^{n} P(A_i \cap B).$$

Therefore,

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(A_i \cap B)}{\sum_{j=1}^{n} P(A_j \cap B)}.$$

This formula can be used to compute all the $P(A_i|B)$ if we are given all the $P(A_i)$ and $P(B|A_i)$.

We can define conditional independence just like independence. We say that A and B are conditionally independent given C if $P(A \cap B|C) = P(A|C)P(B|C)$.

3 Random variables

Definition A random variable is a (measurable) function from the sample space to the real numbers.

Example: $\Omega = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{F}$ = all subsets,

$$X(\omega) = \begin{cases} 1, & \text{if } \omega \in \{2, 4, 6\}, \\ 0, & \text{if } \omega \in \{1, 3, 5\}. \end{cases}$$

Often, the sample space will be implicit and we'll just write X instead of $X(\omega)$.

Examples of discrete random variables:

1. Bernoulli(p): Models the outcome of a coin toss. $P(X = 1) = p$, $P(X = 0) = 1 - p$.

2. Binomial(n, p): Models the number of heads in n coin tosses.

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad 0 \leq k \leq n.$$

   (Is this a valid probability distribution/ probability mass function?) Note: Can construct X as $X = Y_1 + \ldots + Y_n$, where the Y_i are iid Bernoulli(p).

3. Poisson(λ):
$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \ldots$$

4. Geometric(p): Models the number of coin tosses until seeing the first head. $P(X = k) = (1-p)^{k-1}p$, $k = 1, 2, 3, \ldots$

5. Zeta or Zipf distribution: $P(X = k) = Ck^{-(\alpha+1)}$, $k = 1, 2, 3, \ldots$, where $\alpha > 0$ is a specified parameter, and $C > 0$ is a constant chosen so that the probabilities sum to one. In fact, $1/C = \zeta(\alpha + 1)$ where $\zeta(\cdot)$ is the Riemann zeta function, defined for real $s > 1$ as
$$\zeta(s) = 1 + \Big(\frac{1}{2}\Big)^s + \Big(\frac{1}{3}\Big)^s + \cdots$$
It has been proposed as a model for ranked word frequencies in documents (Zipf, 1935), number of species in a genus (Yule, 1924), etc. Suggestion: look up Benford's law.

Where does the Poisson distribution come from? Consider random variables $X_1, X_2, \ldots$ where X_n is Binomial($n, \lambda/n$). Fix $k \geq 0$ and look at $P(X_n = k)$ as n tends to infinity. We have
$$\begin{aligned} P(X_n = k) &= \binom{n}{k} \Big(\frac{\lambda}{n}\Big)^k \Big(1 - \frac{\lambda}{n}\Big)^{n-k} \\ &= \frac{n(n-1)\cdots(n-k+1)}{n \cdot n \cdots n} \frac{1}{k!} \lambda^k \Big(1 - \frac{\lambda}{n}\Big)^n \Big(1 - \frac{\lambda}{n}\Big)^{-k} \\ &\to \frac{1}{k!} \lambda^k e^{-\lambda}. \end{aligned}$$

Roughly speaking, the Poisson distribution models the number of occurences of an event which is *individually rare* but where there is a *large population* of individuals where it could occur. An example is the number of life insurance policy holders of a given age who die in a given year. (This is a bit of a simplification, a compound Poisson would be a better model.) Another example is the number of atoms in a sample undergoing radioactive decay in a given time period. Poisson apparently arrived at this model by studying the number of deaths in the Prussian army due to being kicked by horses.

4 Continuous random variables

In the case of discrete random variables, we were able to specify the probability of each possible outcome. That isn't possible for continuous random variables. What we want is to be able to specify the probability of every "measurable" subset of the real numbers. But I haven't told you what measurable is and, anyway, there are too many such subsets. It turns out that this isn't necessary. It is enough if we specify the probabilities for all intervals of the form $(-\infty, x]$, say.

To make this precise, let X be a random variable. Now, $P(X \in (-\infty, x]) = P(X \leq x)$ is a function of x. Let us denote it by $F(x)$. What are the properties that F must have?

Clearly, F must be non-decreasing and right-continuous (why?), and we must have $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$. The function F is called the distribution function (or cumulative distribution function, cdf) of the random variable X. We write F_X when we want to make it clear which random variable we are talking about.

What probabilities can we calculate from F? We can calculate probabilities for intervals of the form $(x, \infty]$, and of the form $(-\infty, x)$ (see Problem 2.) Hence, we can also compute probabilities of all intervals of the form $[x, y]$ (or open or half-open intervals) where $x \leq y$, and all sets obtained by starting with intervals and performing countably many union and intersection operations. It turns out that this is all we need. These are all the Borel measurable subsets of the real line. Another way of saying this is that intervals of the form $(-\infty, x]$ *generate* the Borel σ-algebra.

Are there subsets of the real line which cannot be generated by the above procedure, i.e., are there sets which are not Borel measurable? The answer is yes, lots of them, but we won't worry about them in this course.

Probability density function: The distribution function F is one way of specifying probabilities for a random variable X. It has the advantage of being completely general, and applying to both discrete and continuous random variables (and mixtures of the two).

Definition: A random variable X is said to be continuous if there is a non-negative function f such that, for any interval (x, y),

$$P(X \in (x, y)) = \int_x^y f(u)du.$$

The function f is called the probability density function of X.

We write f_X if we need to make clear which random variable we are talking about. Observe that $P(X \in (x, y)) = P(x \in [x, y])$ for continuous r.v.s. How is F related to f?

Examples:

1. Uniform($[a, b]$), $a < b$:

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b, \\ 0, & \text{otherwise.} \end{cases}, \quad F(x) = \begin{cases} 0, & x < a, \\ \frac{x-a}{b-a}, & a \le x \le b, \\ 1, & x > b. \end{cases}$$

2. Exponential(λ), $\lambda > 0$: $f(x) = \lambda e^{-\lambda x} 1(x \ge 0)$,

$$F(x) = \begin{cases} 0, & x < 0, \\ 1 - e^{-\lambda x}, & x \ge 0. \end{cases}$$

3. Gamma(α, λ), $\alpha, \lambda > 0$:

$$f(x) = \begin{cases} \frac{1}{\Gamma(\alpha)} \lambda e^{-\lambda x} (\lambda x)^{\alpha - 1}, & x \ge 0 \\ 0, & \text{otherwise,} \end{cases}$$

   where $\Gamma(\alpha) := \int_0^\infty x^{\alpha-1} e^{-x} dx$. Here, α is called the shape parameter and λ is called the scale parameter.

4. Normal(μ, σ^2):

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big).$$

   The parameters μ and σ^2 are in fact the mean and variance of this distribution (to be defined).

The exponential distribution is used to model the lifetime of things whose "frailty" doesn't change with age. What do we mean by this? Let X be an Exp(λ) random variable denoting the lifetime of a light bulb, say. Conditional on the light bulb having survived up to time t, what is the probability that it will survive until time $t + s$? We can calculate this using Bayes' formula. We have

$$\begin{aligned} P(X > t + s | X > t) &= \frac{P(\{X > t + s\} \cap \{X > t\})}{P(X > t)} = \frac{P(X > t + s)}{P(X > t)} \\ &= \frac{\exp(-\lambda(t + s))}{\exp(-\lambda t)} = e^{-\lambda s} = P(X > s). \end{aligned} \tag{2}$$

In other words, the probability that the light bulb will survive for another s time units is the same no matter how old the light bulb is.

Examples of the exponential distribution in nature include the radioactive decay of nuclei, where the probability that a nucleus decays in some time interval $(s, t]$ doesn't depend on how old the nucleus is at time s. (The residual lifetime is independent of the age.)

Of course, this is unusual. Humans, for example, have different probabilities of dying (in a given year, say) at different ages. How do we model this?

Replacing s by an infinitesimal quantity ds in equation (2), we see that, for exponential X, $P(X > t + ds|X > t) = \exp(-\lambda ds) = 1 - \lambda ds$, and so $P(X \leq t+ds|X > t) = \lambda ds$. In other words, the object has a λds probability of dying in the next ds time units, conditional on having been alive up to time t. For this reason, λ is called the hazard rate.

To model a liftetime distribution other than the exponential, we simply replace the constant λ by a non-negative function $\lambda(t)$ of the age t (called the hazard rate function). Thus, $\lambda(t)dt$ denotes the probability of dying during $(t, t + dt]$ conditional on being alive at t. From this, we can work out that the lifetime Y has cumulative distribution function F_Y given by

$$F_Y(t) = \begin{cases} 1 - \exp\Big(-\int_0^t \lambda(s)ds\Big), & t \geq 0, \\ 0, & t < 0. \end{cases}$$

(Check that this lifetime distribution indeed has the claimed hazard rate. What condition on the function $\lambda(\cdot)$ is needed to make it a valid cdf? What does it mean if this condition isn't satisfied?)

Suppose Y is a Gamma random variable with parameter (α, λ) and α is a whole number. Then, we can obtain Y as

$$Y = X_1 + X_2 + \ldots + X_\alpha,$$

where the X_i are iid Exponential random variables with parameter λ.

The normal distribution is possibly the most famous distribution in all of probability. It is also known as the Gaussian distribution, after Carl Friedrich Gauss, who used it to model errors in astronomical observations. It owes its ubiquity in probability and statistics to the Central Limit Theorem, which we'll see later.