Probability – The Highlights

Adam Kelly (`ak2316@cam.ac.uk`)

June 4, 2021

This handout is a crash course which covers the basic ideas of an introductory and not *completely* (but mostly) rigorous probability course. Many, many details will *not* be spelled out (though we won't write anything incorrect), and some 'mathematical extrapolation' will be needed to derive any significant value from this handout. You have been warned – have fun!

1 Setting Up Probability

We are going to build up probability from the very basic axioms. This will give us a firm mathematical basis to work from, but also will match the intuitive notions of probability that you have hopefully developed before reading this handout.

1.1 Probability Spaces

The basic object of study in probability theory is that of *probability spaces.* We begin by defining how we can talk about the idea of 'events' formally. What we want from our definition will be some way to have a set of possible 'outcomes' (a *sample space*), from which we can combine different outcomes to form various 'events'. This leads almost directly to the definition of an *event space.*

Definition 1.1 (Event Space)**.** The collection $\mathcal{F}$ of subsets of the sample space Ω is an **event space** or **σ-algebra** if

1. $\mathcal{F}$ is non-empty,
2. if $A \in \mathcal{F}$, then $\Omega \backslash A \in \mathcal{F}$,
3. if $A_1, A_2, \dots \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Within this system, $A \cup B$ corresponds to either A or B occurring, $A \cap B$ corresponds to both A and B occurring, and $\Omega \backslash A$ corresponds with anything but A occurring. You get the idea.

Our main goal in the study of probability is to assign some sort of 'likelihood' to events in the event space. This is done through the introduction of a *probability measure.*

Definition 1.2 (Probability Measure)**.** We say that a function $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ is a **probability measure** on $(\Omega, \mathcal{F})$ if

1. $\mathbb{P}(A) \geq 0$ for all $A \in \mathcal{F}$,
2. $\mathbb{P}(\Omega) = 1$,
3. if $A_1, A_2, \dots$ are disjoint events in $\mathcal{F}$, then $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

These properties of the probability measure should match your own intuition about the way that probability works. Putting all the pieces together, we get our main object of study: the *probability space*.

Definition 1.3 (Probability Space)**.** A **probability space** is a triple $(\Omega, \mathcal{F}, \mathbb{P})$ such that Ω is non empty, $\mathcal{F}$ is an event space on Ω, and $\mathbb{P}$ is a probability measure on $(\Omega, \mathcal{F})$.

If we only have to deal with say finite (or even countable) Ω, then it makes sense to usually take $\mathcal{F}$ to just be the power set of Ω. In this case, if $\Omega = \{\omega_1, \omega_2, \dots\}$, then we can construct any probability measure $\mathbb{P}$ by just specifying $\mathbb{P}(\{\omega_i\})$ for all i, as then for $A \in \mathcal{F}$, we can get $\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\{\omega\})$.

1.2 Conditional Probability and Independence

In our day-to-day experience, we know that knowledge about one event can influence how likely we determine another event is to occur. This type of thinking motivates the introduction of *conditional probability.*

Definition 1.4 (Conditional Probability)**.** If $A, B \in \mathcal{F}$ and $\mathbb{P}(B) > 0$, the **probability of A given B** is denoted $\mathbb{P}(A \mid B)$, and is given by

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

It's not hard to check that $\mathbb{P}(\,\cdot \mid B)$ is a probability measure, as we would expect – it's just assigning probabilities to events in $\mathcal{F}$ in a different way!

A natural thing to care about is if knowledge about an event *doesn't* influence how likely another event is. We call such events independent.

Definition 1.5 (Independent Events)**.** We say that events A, B in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ are **independent** if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B),$$

and **dependent** otherwise.

We can see immediately that A, B independent implies that

$$\mathbb{P}(A \mid B) = \mathbb{P}(A) \quad \text{and} \quad \mathbb{P}(B \mid A) = \mathbb{P}(B),$$

if $\mathbb{P}(A), \mathbb{P}(B) \neq 0$. However, this definition is slightly more general than this property since it handles events with zero probability. We can generalise this notion of independence to multiple events, and we get a criterion that's stronger than pairwise independence[1].

Definition 1.6 (Mutually Independent Events)**.** We say that the events A_1, A_2, A_3, ... are **mutually independent** if

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_r}) = \mathbb{P}(A_{i_1})\mathbb{P}(A_{i_2}) \cdots \mathbb{P}(A_{i_r}),$$

for any $i_1, i_2, \dots, i_r$ and $r \geq 2$.

[1] It's not hard to construct an example of events that are pairwise independent but wouldn't be independent in any reasonable sense.

So if conditional probability tells us about the likelihood of a certain event in one particular scenario, then it's easy to imagine that knowing about the likelihood in every scenario would then just give you the likelihood in general. Such thinking is formalised in the *law of total probability*, also known as the *partition theorem*.

Theorem 1.7 (Partition Theorem)**.** *Let $\{B_1, B_2, \dots\}$ be a collection of disjoint events such that $\mathbb{P}(B_i) > 0$ for all i and $\bigcup_{i=1}^{\infty} B_i = \Omega$. Then*

$$\mathbb{P}(A) = \sum_i \mathbb{P}(A \mid B_i)\mathbb{P}(B_i),$$

for any $A \in \mathcal{F}$.

Proof. We can compute that

$$\sum_i \mathbb{P}(A \mid B_i)\mathbb{P}(B_i) = \sum_i \mathbb{P}(A \cap B_i) = \mathbb{P}\left(\bigcup_i (A \cap B_i)\right) = \mathbb{P}(A),$$

as required. □

Using the partition theorem, we can establish the famous **Bayes' theorem**, which allows us to use knowledge about $\mathbb{P}(A \mid B_j)$ to give information about $\mathbb{P}(B_j \mid A)$:

$$\mathbb{P}(B_j \mid A) = \frac{\mathbb{P}(A \mid B_j)\mathbb{P}(B_j)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A \mid B_j)\mathbb{P}(B_j)}{\sum_i \mathbb{P}(A \mid B_i)\mathbb{P}(B_i)}.$$

1.3 Continuity of Probability Measures

In the definitions for event spaces (σ-algebras) and probability measures, you may notice that we specify rules for countably infinite sequences of events. This is for good reason – there's plenty of natural scenarios where such a setup is useful. For example, consider the experiment of tossing a coin until you get heads, in which it is possible to get an arbitrarily long run of tails before you stop tossing. To make working with such scenarios easier, we can establish a result which allows us study events as limits of a sequence of events – that the probability measure is continuous.

Definition 1.8 (Limits of Events)**.** We say that a sequence of events $A_1, A_2, \dots$ is *increasing* if $A_1 \subseteq A_2 \subseteq \cdots$. We define the **limit** as

$$\lim_{n \to \infty} A_n = \bigcup_{n=1}^{\infty} A_n.$$

Theorem 1.9 (Continuity of Probability Measures)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. If $A_1, A_2, \dots$ is an increasing sequence of events in $\mathcal{F}$ with limit A, then $\mathbb{P}(A) = \lim_{n\to\infty} \mathbb{P}(A_n)$.*

Proof. Define $B_i = A_i \backslash A_{i-1}$, with $B_1 = A_1$. Then we have

$$A = B_1 \cup B_2 \cup B_3 \cup \cdots$$

which is the union of disjoint events in $\mathcal{F}$. Thus we have

$$\begin{aligned}\mathbb{P}(A) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) &= \mathbb{P}\left(\bigcup_{n=1}^{\infty} B_n\right)\\ &= \lim_{n\to\infty}\sum_{i=1}^{n}\mathbb{P}(B_n) = \lim_{n\to\infty}\mathbb{P}(B_1\cup\cdots\cup B_n)\\ &= \lim_{n\to\infty}\mathbb{P}(A_n),\end{aligned}$$

as required. □

As a side note, the type of construction used to define B_i can be quite helpful when we want to deal with a disjoint union, rather than some non-disjoint sequence of events (since it works nicely with the definition of the probability space!)

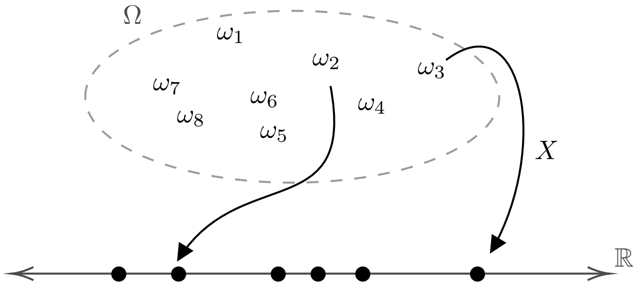
2 Discrete Random Variables

Imagine the following scenario. You are in a lecture hall, and you decide to perform an experiment where choose a student at random and ask them how they performed in the admissions test. It would be natural to model this scenario with a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ where $\Omega = \{\omega_1, \dots, \omega_n\}$ where each ω_i corresponds to a student in the lecture hall, $\mathcal{F}$ is all possible subsets of Ω, and $\mathbb{P}$ is the probability measure describing the likelihoods of given students being chosen in the experiment.

So how does the 'asking for the student's test score' fit into this setup? A natural way to deal with this is by defining a function $X : \Omega \to \mathbb{R}$ where $X(\omega_i)$ is the admissions test score of the student ω_i. We call such a function a *discrete random variable.*

Definition 2.1 (Discrete Random Variables)**.** A **discrete random variable** X on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a function $X : \Omega \to \mathbb{R}$ such that the image $X(\Omega)$ is a countable subset of $\mathbb{R}$ and $\{\omega \in \Omega : X(\omega) = x\} \in \mathcal{F}$ for $x \in \mathbb{R}$.

A natural mental picture to go along with this definition (which roughly follows our motivating example) is shown below.



2.1 Probability Mass Functions

With the introduction of discrete random variables, we can change our view from probabilities of events occurring to the probabilities that a discrete random variable takes on specific values. A nice way to do this is by introducing the *probability mass function.*

Definition 2.2 (Probability Mass Function)**.** The **probability mass function** or **pmf** of a discrete random variable X is a function $p_X : \mathbb{R} \to [0,1]$ defined by

$$p_X(x) = \mathbb{P}(X = x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) = x\}).$$

It turns out that the probability mass function encodes pretty much all information about the discrete random variable X, so much so that if we have some function p_X that acts like a probability mass function, we can say things like 'let X be a random variable taking value x_i with probability $p_X(x_i)$' and have it be completely well defined.

Theorem 2.3. *Let $p_X : \mathbb{R} \to [0,1]$ be a function such that $p_X(x) \neq 0$ for countably many x and $\sum_{x\in\mathbb{R}} p_X(x) = 1$. Then there is a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a random variable X on this space where X has probability mass function p_X.*

Proof. Take $\Omega = \{x \in \mathbb{R} : p_X(x) \neq 0\}$, $\mathcal{F}$ to be all subsets of Ω, and

$$\mathbb{P}(A) = \sum_{\omega\in A} p_X(\omega) \quad \text{for } A \in \mathcal{F},$$

and define $X : \Omega \to \mathbb{R}$ by $X(\omega) = \omega$. □

2.2 Expected Value

A common question when dealing with discrete random variables is 'what is the mean of this discrete random variable?'. Thinking back to our lecture hall example, this would correspond with asking what the mean test score in the lecture hall was. We answer such questions by defining the *expectation* of a random variable.

Definition 2.4 (Expected Value)**.** If X is a discrete random variable, we define the **expected value** of X to be

$$\mathbb{E}[X] = \sum_{x\in\operatorname{img} X} x\mathbb{P}(X = x)$$

whenever this sum converges absolutely[2].

There are a few 'tricks' which make the computation of expected value somewhat easier. For example, if we have a discrete random variable defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, then we can sum over the elements of Ω rather than the elements of $\operatorname{img} X$. Also, since random variables are really just functions, we can obtain new random variables by composing them with other functions, for example $X \mapsto X^2$. In such scenarios, we can use a seemingly obvious result.

[2] We require this because we rearranging the sum shouldn't change the expected value – see Analysis I.

Theorem 2.5 (Law of The Subconscious Statistician)**.** *If X is a discrete random variable and $g : \mathbb{R} \to \mathbb{R}$, then*

$$\mathbb{E}[g(X)] = \sum_{x \in \operatorname{img} X} g(x)\mathbb{P}(X = x),$$

whenever this sum converges absolutely.

Proof. Let $Y = g(X)$ be a discrete random variable. Then

$$\begin{aligned}\mathbb{E}[Y] &= \sum_{y \in \operatorname{img} Y} y\mathbb{P}(Y = y) = \sum_{y \in \operatorname{img} Y} y \sum_{x \in \operatorname{img} X : g(x) = y} \mathbb{P}(X = x) \\ &= \sum_{x \in \operatorname{img} X} g(x)\mathbb{P}(X = x),\end{aligned}$$

provided this converges absolutely. □

Another thing we can do very easily is find the expected value of a sum of random variables, using *linearity of expectation.*

Theorem 2.6 (Linearity of Expectation)**.** *For any random variables $X_1, X_2, \ldots, X_n$ for which the following expectations exist,*

$$\mathbb{E}\left[X_1 + \cdots + X_n\right] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n].$$

Proof. We sum over the elements of the sample space Ω to get

$$\begin{aligned}\mathbb{E}[X] &= \sum_{\omega \in \Omega} p(\omega)[X_1(\omega) + \cdots + X_n(\omega)] \\ &= \sum_{\omega \in \Omega} p(\omega)X_1(\omega) + \cdots + \sum_{\omega \in \Omega} p(\omega)X_n(\omega) \\ &= \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n].\end{aligned}$$

□

It's important to note that there's absolutely no requirement that the random variables be independent for this to hold – one of the reasons why it's an incredibly useful result!

So if expected value measures something like the 'mean' of a random variable, it's natural to think about how we can measure how we can expect the random variable deviates from this mean. This is done through the notion of *variance.*

Definition 2.7 (Variance)**.** The **variance** $\operatorname{var}(X)$ of a discrete random variable X is defined to be

$$\operatorname{var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

By just working through the definitions we can see that $\operatorname{var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$, which is often an easy way to calculate this quantity.

2.3 Indicator Functions

There's one type of random variable which is so useful that it needs to be brought to your attention immediately – indicator functions.

Definition 2.8 (Indicator Functions)**.** The **indicator function** of an event A is the random variable $\mathbb{1}_A$ where

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

Indicator functions have nice algebraic facts which reflect common ways to manipulate events:

$$\begin{aligned} \mathbb{1}_{A^c} &= 1 - \mathbb{1}_A, \\ \mathbb{1}_{A\cap B} &= \mathbb{1}_A \mathbb{1}_B, \\ \mathbb{1}_{A\cup B} &= \mathbb{1}_A + \mathbb{1}_B - \mathbb{1}_A \mathbb{1}_B. \end{aligned}$$

Still one of the best things about indicator functions is that they allow us to talk about probabilities of events using expectations, since

$$\mathbb{E}[\mathbb{1}_A] = \mathbb{P}(A).$$

All of these properties are straightforward and easy to check (just do casework), but they can make proving things and solving problems much easier. Nontrivial example time!

Problem 2.9. A total of n bar magnets are placed end to end in a line with random independent orientations. Adjacent like poles repel, ends with opposite polarities join to form blocks. Let X be the number of blocks of joined magnets. Find $\mathbb{E}[X]$ and $\operatorname{var}(X)$.