Primer on Random Variables

S. Simmons

December 5, 2021

Contents

1 Preliminaries

Probability theory is concerned with certain non-deterministic processes (e.g., flipping a coin, rolling a pair of dice, and computing some sample statistic after taking a simple random sample from a given population) whose outcome cannot be predicted with certainty. These so-called *random* experiments can be repeated, and each trial produces an outcome from a well-defined set of possibilities. But the particular outcome of a trial can be determined only by the performing the experiment and then observing the result.

Random variables are function-theoretic constructs that allow us to factor outcomes of experiments into components. When rolling at once 5 dice, for example, we might use a model with one random variable for each die. Moreover, in practice, random variables suppress certain inessential technical details in favor of facts salient to the computations we wish to make. Mathematically, a random variable is a *push-forward.* We will explain this below via specific examples. First, what is the definition of a random variable?

give a simple example here??

A random variable, $\mathcal{X}$, is a map $\mathcal{X} : \Omega \to \mathbb{R}$ where Ω is the sample space for a probability space. But what is a probability space? The concept of a probability space is central to the modern axiomatic development (a framework due to Kolmogorov) of probability theory. Before diving into proper thinking about random variables, let us unpack the definition of a probability space, along with its motivation.

def:probspace

Definition 1. A *probability space* is a triple $(\Omega, \mathcal{F}, P)$ where Ω is a set of outcomes, $\mathcal{F}$ is collection of subsets of Ω (that is a σ-algebra; i.e., closed under countable set-theoretic operations), and P is a probability measure on $\mathcal{F}$ (that is σ-additive).

Think of the sample space[1] Ω in the definition above as the set of possible outcomes of a random experiment; for example, rolling a single six-sided die is a random experiment for which $\Omega = \{1, 2, 3, 4, 5, 6\}$. We would like to define a function (cf. the definition of *probability mass function* below) $f : \Omega \to [0, 1]$ whose output $f(\omega)$ is the probability of the occurrence of outcome ω, and that satisfies $\sum_{\omega \in \Omega} f(\omega) = 1$. For a fair six-sided die, we set $f(\omega) = 1/6$ for each $\omega \in \{1, 2, 3, 4, 5, 6\}$. Note that, for the existence of such a function f, the sample space need not be finite; e.g., if $\Omega = \mathbb{N}$, the positive integers, then one may set $f(n) = 1/2^n$.

But what if the experiment is drawing a real number at random from the interval $\Omega = [0, 1]$ (or drawing an element from any sample space that is uncountably infinite)? If Ω is uncountable, we cannot[2] assign values $f(\omega) > 0$ on Ω (or on any uncountable subset of Ω) in such way that $\sum_{\omega \in \Omega} f(\omega) = 1$.

is $\mathcal{F}$ really a class

A way around this problem is to introduce a set $\mathcal{F}$ consisting of certain subsets (which we describe in detail below) of Ω. Then, instead of specifying probabilities for every individual outcome ω in Ω, we define a function $P : \mathcal{F} \to [0, 1]$ called a *probability measure* in such a way that, for $A \in \mathcal{F}$, $P(A)$ is the *probability* (or likelihood) of A occurring. An element $A \in \mathcal{F}$ (i.e., a subset of Ω) is called an *event*, and $P(A) \in [0, 1]$ is the probability that the event A occurs in a trial of the experiment (here, the event A *occurring* means that $\omega \in A$ where ω is the outcome of the trial). When rolling a six-sided die, we might be interested in $P(\{2, 4, 6\})$, the likelihood that a roll of the die turns up even.

A probability space for which Ω is finite or countably infinite is called *discrete*. Basic experiments involving dice or cards, for example, often involve discrete probability spaces. Let us introduce two somewhat more involved examples. In computer science, we might wish to simulate a random process that selects a real number from the interval $[0, 10]$, and that does so *uniformly*, meaning that every number has equal chance of being selected. In this case, the sample space Ω is the (mathematically, at least) uncountable set $[0, 10]$. We might be interested in the event that the real number chosen is in $[1, 1.5] \cup [8.5, 10]$.

Below, we consider the process of continually flipping a coin. For this experiment, the sample space Ω consists of all countably infinite strings built from the characters T and H. Note that here, too, Ω is uncountable, as can be shown using a Cantor diagonalization argument. In this setting, we might be interested in the probability that the first *heads* occurs on the fourth flip of the coin, or that six flips in row are *heads*.

Returning to our discussion of probability spaces: the set $\mathcal{F}$ in Definition 1 consists then of events A (meaning: subsets of Ω) to which we can directly apply the probability measure P in order to compute the corresponding probability $P(A)$. Jumping ahead to the endgame: given a sample space, Ω, we want to define $\mathcal{F}$ and P in such a way that $\mathcal{F}$ contains *every* event for which we can *consistently* use P to compute its probability of occurrence. This is where the algebra comes in.

[1] The sample space Ω is just a set; yet, as is common in the literature, we often refer to Ω as a "space". Below we will state the technically correct definition which is that a sample space is a pair $(\Omega, \mathcal{F})$ where Ω is a set and $\mathcal{F}$ is σ-algebra of subsets of Ω.

[2] From Calculus we know how to use partial sums to sum a countable number of numbers; but can we even define an uncountable sum? In the case where $f(x) > 0$ for all x in an uncountable set Ω, even if we try to set $\sum_{x \in \Omega} f(x) = \sup_{I \subset \Omega}\{\sum_{x \in I} f(x)\}$ where the supremum is over all countable subsets I of Ω, we have that $\Omega = \bigcup_{n=1}^{\infty} A_n$ where $A_n = \{x \in \Omega \mid f(x) > 1/n\}$; hence at least one of the sets A_n must be uncountable since otherwise Ω is countable, it being a countable union of countable sets. But a sum involving an infinite number of numbers greater than a fixed positive number $1/n$ cannot be 1 (or even a finite number).

1.1 Algebras of sets

In the six-sided die experiment, in which $\Omega = \{1, 2, 3, 4, 5, 6\}$, we can simply take $\mathcal{F}$ to be the power set (the set of all subsets) of Ω. In the case of a *fair* six-sided die, we then specify the probability measure P by declaring that $P(\emptyset) = 0$ and $P(\Omega) = 1$ and requiring the basic property familiar for elementary probability theory: that for events A_1 and A_2 that are disjoint (meaning that $A_1 \cap A_2 = \emptyset$), we must have that $P(A_1 \cup A_2) = P(A_1) + P(A_2)$. This then completely determines (for a fair die) P on all of $\mathcal{F}$; e.g., $P(\omega) = 1/6$ for each ω in Ω, and $P(\{2, 4, 6\}) = 1/2$.

Taking $\mathcal{F}$ to be the power set of Ω will not work in the case of the random number simulator or the continual coin flipping experiment (at least not in the natural interpretation of those in which P is almost never zero): we immediately run into trouble with the normalization $\sum_{\omega\in\Omega} P(\omega) = 1$ since Ω is uncountable. The set $\mathcal{F}$ being large is good since then lots of events have associated probabilities, but it cannot be too large.

In cases in which we cannot simply take $\mathcal{F}$ to be the power set of Ω, exactly how large should $\mathcal{F}$ be? With the stipulation mentioned above that we want $\mathcal{F}$ to contain not just some, but all, *measureable* events, we expect the events in $\mathcal{F}$ to be not only large in number but also quite varied. In a modern axiomatic development of probability theory we want to get our hands on these sets easily and efficiently. How is $\mathcal{F}$ defined in the literature and how do we conceptualize its definition in terms of first principles?

Clearly we should be able to compute the probability the complement of a given event as well as that of the unions and intersections of events:

Definition 2. Given a set Ω, an *algebra*[3] *(of sets)* Ω is a collection $\mathcal{A}$ of subsets of Ω that satisfies

(i) $\Omega \in \mathcal{A}$;

(ii) closure under complements: $A \in \mathcal{A}$ implies that the complement, $\Omega \setminus A$, is again in $\mathcal{A}$; and

(iii) closure under finite unions: $A_1, A_2, \ldots, A_n \in \mathcal{A}$ implies that $A = \bigcup_{i=1}^{n} A_i \in \mathcal{A}$.

From parts (i) and (ii) we immediately see that any algebra (of sets of some sample space) contains the empty set. In fact, $\{\emptyset,\ \Omega\}$ the smallest algebra possible for a given Ω. The largest possible algebra is the power set of Ω, which as we noted above could be too large for our purposes.

Example 1. Let $\Omega = \mathbb{R}$. The collection of all finite unions of sets of the form $[a, b) \coloneqq \{x \in \mathbb{R} \mid a \le x < b\}$ where $-\infty \le a < b \le \infty$ is an algebra.

As far as intersections, we have part (i) of the following proposition, which follows easily from the appropriate De Morgan's law. (We could have assumed closure under finite intersections in Definition 2 and then used the other De Morgan's law to prove closure under finite unions in Proposition 2.)

Proposition 1. *Let $\mathcal{A}$ be an algebra (of subsets of a sample space Ω).*

(i) $\mathcal{A}$ is closed under finite intersections: if $A_1, A_2, \ldots, A_n \in \mathcal{A}$, then $\bigcap_{i=1}^{n} A_i \in \mathcal{A}$.

[3] Arithmetic on the level of sets models that of numbers: union and intersection are analogous to addition and multiplication; complementation is somewhat analogous to negation; and $\le$ corresponds to $\subset$. In the literature, some authors use the term *field (of sets)* instead of *algebra*.

(ii) If $A_1, A_2 \in \mathcal{A}$, then $A_1 \setminus A_2 \in \mathcal{A}$.

Proof. Let $A_1, A_2, \dots, A_n$ be sets in $\mathcal{A}$. Combining the facts (from Definition 2) that an algebra is closed under both complements and finite unions, we have that $\bigcup_{i=1}^{n}(\Omega \setminus A_i) \in \mathcal{A}$. But, by De Morgan's law, $\Omega \setminus (\bigcap_{i=1}^{n} A_i) = \bigcup_{i=1}^{n}(\Omega \setminus A_i)$, so that $\Omega \setminus (\bigcap_{i=1}^{n} A_i) \in \mathcal{A}$ and, taking the complement again, $\bigcap_{i=1}^{n} A_i = \Omega \setminus (\Omega \setminus (\bigcap_{i=1}^{n} A_i)) \in \mathcal{A}$, proving part (i). Next let A_1 and A_2 be sets in $\mathcal{A}$. Then $A_1 \setminus A_2 = A_1 \cap (\Omega \setminus A_2) \in \mathcal{A}$, by part (i), □

We can think of the algebra in Example 1 as being *generated* by the intervals of the form $[a, b)$. In fact, if $\mathcal{S}$ is any collection of subsets of some sample space Ω, we can define the algebra generated by $\mathcal{S}$ to be the smallest[4] algebra of sets in Ω that contains $\mathcal{S}$. Equivalently, said smallest algebra can be generated as follows. Let $\mathcal{S}_1$ to be the set consisting of all elements of $\mathcal{S}$ along with their complements: $\mathcal{S}_1 = \mathcal{S} \cup \{\Omega \setminus S \mid S \in \mathcal{S}\}$. Next define $\mathcal{S}_2$ as the set of all finite intersections of elements of $\mathcal{S}_1$. Then the algebra generated by $\mathcal{S}$ is the set consisting of all finite unions of element in $\mathcal{S}_2$. (Note that in Example 1, $\mathcal{S}_2 = \mathcal{S}_1 = \mathcal{S} = \{[a, b) \mid -\infty \leq a < b < \infty\}$ so that the algebra generated by $\mathcal{S}$ is indeed the set of all finite unions of intervals of the form $[a, b)$.)

1.2 Probability measures

Let $\mathcal{A}$ be an algebra of sets in a some sample space Ω. Think of an element $A \in \mathcal{A}$ as an event to which we want to assign a probability $P(A)$. How should P be defined?

A *partition* of Ω is a collection of pairwise-disjoint subsets of Ω whose union is Ω. From basic probability theory — thinking of the probability $P(A)$ of an event A as the proportion of times, over many trials, that the event A occurs — we know that if that the events $A_1, A_2, \dots, A_n \in \mathcal{A}$ partition the sample space Ω of a probability space then, since one and only one A_i occurs in a trial, we must require that $P(A_1) + P(A_2) + \cdots + P(A_n) = 1$. Note that this implies, as a degenerative special case, that $P(\Omega) = 1$.

Now suppose that there exists a countably infinite partition of Ω: events $A_i \in \mathcal{A}$, $i \in \mathbb{N}$ satisfying $\bigcup_{i=1}^{\infty} A_i = \Omega$ and $A_i \cap A_j = \emptyset$ when $i \neq j$. From basic considerations, we know only that we cannot not have $\sum_{i=1}^{\infty} P(A_i) > 1$; hence, for a countably infinite partition of a sample space, we might consider requiring that P satisfy $\sum_{i=1}^{\infty} P(A_i) \leq 1$.

But suppose that we allow that the inequality possibly be strict, $\sum_{i=1}^{\infty} P(A_i) < 1$, and consider the sets $B_n = \Omega \setminus \bigcup_{i=1}^{n} A_i = \bigcup_{i=n+1}^{\infty} A_i$. Each B_i (being the complement of a finite union of sets in $\mathcal{A}$) is itself in $\mathcal{A}$ (since $\mathcal{A}$ is an algebra); so that we now have a non-increasing sequence of events, $B_0 := \Omega \supset B_1 \supset B_2 \supset \cdots$ which converges[5] to the empty set (since the A_i partition Ω). Intuitively, we want that $\lim_{n\to\infty} P(B_n) = 0$. However, since the A_i are mutually disjoint, we have, for each $n \in \mathbb{N}$, that the events $B_n, A_1, A_2, \dots, A_n$ are mutually disjoint, and in fact form a finite partition of Ω. Hence, for each n, $P(B_n) + \bigcup_{i=1}^{n} P(A_i) = 1$ or, equivalently, $P(B_n) = 1 - \bigcup_{i=1}^{n} P(A_i)$. Now, taking limits, we see that if $\sum_{i=1}^{\infty} P(A_i) < 1$, then there is no way that $\lim_{n\to\infty} P(B_n) = 0$ since $\lim_{n\to\infty} P(B_n) = \lim_{n\to\infty} 1 - \sum_{i=1}^{n} P(A_i) = 1 - \sum_{i=1}^{\infty} P(A_i) > 0$.

and for
. . . reasons . . .

[4] If we take the intersection of a collection of algebras, each of which is an algebra of sets for the same sample space Ω, then we again get an algebra of sets in Ω. Since there exists at least one algebra containing $\mathcal{S}$ (e.g., the power set of Ω), we can define the *smallest* algebra for Ω containing $\mathcal{S}$ to be the intersection of all algebras for Ω that contain $\mathcal{S}$. If $\mathcal{A}_{\mathcal{S}}$ denotes this intersection, then $\mathcal{A}_{\mathcal{S}}$ is minimal in that $\mathcal{A}_{\mathcal{S}} \subset \mathcal{A}$ for any algebra $\mathcal{A}$ containing $\mathcal{S}$.

[5] A sequence of sets $A_1, A_2, \dots$ is *non-decreasing* if $A_1 \subset A_2 \subset \cdots$; it is *non-increasing* if $A_1 \supset A_2 \supset \cdots$. If a sequence $\{A_i\}_{i=1}^{\infty}$ is non-decreasing, we define $\lim_{i\to\infty} A_i = \bigcup_{i=1}^{\infty} A_i$; if it is non-increasing, we define $\lim_{i\to\infty} A_i = \bigcap_{i=1}^{\infty} A_i$. A sequence of sets is *monotonic* if it is either non-decreasing or non-increasing.

In light of the previous paragraph, one formulates the following definition.

Definition 3. Let $\mathcal{A}$ be an algebra of sets in Ω. A function $P : \mathcal{A} \to \mathbb{R}$ is a *probability measure* if

(i) $P(A) \geq 0$ for all $A \in \mathcal{A}$, and

(ii) $\sum_{i=1}^{\infty} P(A_i) = 1$ whenever $\{A_i\}_{i=1}^{\infty}$ partitions Ω; i.e., whenever $A_i \in \mathcal{A}$ for all i, $\bigcup_{i=1}^{\infty} A_i = \Omega$, and the A_i are mutually disjoint.

Using only the definition of an algebra and the definition of a probability measure, we easily prove the following.

Proposition 2. *Let $\mathcal{A}$ be an algebra of sets in Ω and let P be a probability measure on $\mathcal{A}$. Then*

(i) $P(\emptyset) = 0,$

(ii) $P(A) \in [0, 1]$ *for all* $A \in \mathcal{A},$

(iii) $P(\Omega) = 1,$

(iv) $\sum_{i=1}^{n} P(A_i) = 1$ *for any finite partition* $A_1, A_2, \ldots, A_n$ *of* Ω *by events* $A_i \in \mathcal{A}$*, and*

(v) $P(\bigcup_i A_i) = \sum_i P(A_i)$ *for any countable (finite or countably infinite) family* $\{A_i\}$ *of pairwise disjoint events in* $\mathcal{A}$ *for which* $\bigcup_i A_i$ *is in* $\mathcal{A}$ *(which is automatically the case for a finite family since* $\mathcal{A}$ *is an algebra).*

Proof. Let $\{A_i\}_{i=1}^{\infty}$ be a partition of Ω (for example, one can take $A_1 = \Omega$ and $A_i = \emptyset$ for $i > 1$). Then, $\{\emptyset, A_1, A_2, \ldots\}$ also being a partition of Ω, we have both $\sum_{i=1}^{\infty} P(A_i) = 1$ and $P(\emptyset) + \sum_{i=1}^{\infty} P(A_i) = 1$; hence, $P(\emptyset) = 0$, proving (i).

Part (iv) now follows from Definition 3, part (ii) since we can view a finite partition of Ω as a countably infinite one by throwing in countably infinitely many copies of the empty set.

To prove (ii), let $A \in \mathcal{A}$. Then $\Omega \setminus A \in \mathcal{A}$ and

$$P(A) + P(\Omega \setminus A) = 1 \tag{1}$$

by (iv) since $\{A,\ \Omega \setminus A\}$ is a partition of Ω by events in $\mathcal{A}$. Both terms on the left of (1) being non-negative implies that $0 \leq P(A) \leq 1$.

Taking $A = \Omega$ in (1), we obtain part (iii) as a consequence of (i).

Lastly, let $\{A_i\}$ be a countable family of sets in $\mathcal{A}$ and suppose that $\bigcup_i A_i$ is also in $\mathcal{A}$. Then $A := \Omega \setminus \bigcup_i A_i$ is in $\mathcal{A}$ and both $\{A, A_1, A_2, \ldots\}$ and $\{A,\ \bigcup_i A_i\}$ are partitions of Ω. Hence, $P(A) + \sum_i P(A_i) = 1 = P(A) + P(\bigcup_i A_i)$ and, subtracting $P(A)$ on the left and right, we obtain (v). □

The reader may have noticed that Definition 3 and Proposition 2 can be reorganized. In fact, one can prove that, if $\mathcal{A}$ is an algebra for Ω, then $P : \mathcal{A} \to \mathbb{R}$ being a probability measure is equivalent to parts *(ii)*, *(iii)*, and *(v)* of Proposition 2 all holding for P.

The statements in the following proposition are familiar from basic probability theory and can readily be proved.

Proposition 3. *Let $\mathcal{A}$ be an algebra for Ω.*

(i) If $A \in \mathcal{A}$, then $P(\Omega \setminus A) = 1 - P(A)$.

(ii) If $A_1, A_2 \in \mathcal{A}$, then $P(A_1 \cup A_2) = P(A_1) + P(A_1) - P(A_1 \cap A_2)$.

Let us check that we have the desirable behavior regarding limits.

Proposition 4. *Let $\{A_i\}_{i=1}^{\infty}$ be a monotonic sequence of events in an algebra $\mathcal{A}$ of subsets of Ω.*

(i) If $\{A_i\}_{i=1}^{\infty}$ is non-decreasing, then $\lim_{i\to\infty} P(A_i) = P(\bigcup_{i=1}^{\infty} A_i)$.

(ii) If $\{A_i\}_{i=1}^{\infty}$ is non-increasing, then $\lim_{i\to\infty} P(A_i) = P(\bigcap_{i=1}^{\infty} A_i)$.

Proof. Suppose that $\{A_i\}_{i=1}^{\infty}$ is a non-decreasing sequence of events in $\mathcal{A}$: $A_1 \subset A_2 \subset \cdots$. Set $A_0 = \emptyset$ and construct the following countable sequence of pairwise disjoint events (which are also in $\mathcal{A}$, by Proposition 1):

$$A_1 \setminus A_0,\ A_2 \setminus A_1,\ A_3 \setminus A_2,\ \ldots.$$

Note that we have the equality (of sets):

$$\bigcup_{i=1}^{\infty} (A_i \setminus A_{i-1}) = \bigcup_{i=1}^{\infty} A_i;$$

so that, applying Proposition 2, part (v), we have that

$$P(\textstyle\bigcup_{i=1}^{\infty} A_i) = P(\bigcup_{i=1}^{\infty}(A_i \setminus A_{i-1})) = \sum_{i=1}^{\infty} P(A_i \setminus A_{i-1}). \tag{2}$$

Now, applying Proposition 2 part (v) again, we obtain

$$\begin{aligned}\lim_{n\to\infty} P(A_n) &= \lim_{n\to\infty} P(A_1 \setminus A_0 \cup A_2 \setminus A_1 \cup \cdots \cup A_n \setminus A_{n-1}) \\ &= \lim_{n\to\infty} P(A_1 \setminus A_0) + P(A_2 \setminus A_1) + \cdots + P(A_n \setminus A_{n-1}) \\ &= \textstyle\lim_{n\to\infty} \sum_{i=1}^{n} P(A_i \setminus A_{i-1}) \qquad (3)\end{aligned}$$

The terms of the sequence in (3) are just the partial sums of the series on the right in (2). But the sequence is a monotonic non-decreasing sequence (of non-negative real numbers) that is bounded above by 1, so it converges; i.e., $\lim_{n\to\infty} P(A_n) = \lim_{n\to\infty} \sum_{i=1}^{n} P(A_i \setminus A_{i-1}) = \sum_{i=1}^{\infty} P(A_i \setminus A_{i-1}) = P(\bigcup_{i=1}^{\infty} A_i)$, which is (i).

Now suppose that $\{A_i\}_{i=1}^{\infty}$ is a non-increasing sequence of events in $\mathcal{A}$; that is, $A_1 \supset A_2 \supset \cdots$. Then $\{\Omega \setminus A_i\}_{i=1}^{\infty}$ is a non-decreasing sequence to which we can apply (i), obtaining $\lim_{i\to\infty} P(\Omega \setminus A_i) = P(\bigcup_{i=1}^{\infty} \Omega \setminus A_i)$.

By the appropriate De Morgan's law, we know that $\bigcup_{i=1}^{\infty} \Omega \setminus A_i = \Omega \setminus (\bigcap_{i=1}^{\infty} A_i)$ so that $P(\bigcap_{i=1}^{\infty} A_i) = 1 - P(\Omega \setminus (\bigcap_{i=1}^{\infty} A_i)) = 1 - P(\bigcup_{i=1}^{\infty} \Omega \setminus A_i) = 1 - \lim_{i\to\infty} P(\Omega \setminus A_i) = 1 - (\lim_{i\to\infty} 1 - P(A_i)) = \lim_{i\to\infty} P(A_i)$ which proves (ii). □

1.3 σ-algebras

So far, we have shown that if a monotone sequence of events in an algebra $\mathcal{A}$ has a limit that happens to again be in $\mathcal{A}$, then the probability of the limit of the events is the limit of the probabilities of the events — a property the we need for various convergence issues below. verify

But the limit of a monotone sequence of events in an algebra $\mathcal{A}$ need not be in $\mathcal{A}$. For instance, for the algebra $\mathcal{A}$ in Example 1, $\lim_{n=1}^{\infty}[0, 1/n) = \bigcap_{n=1}^{\infty}[0, 1/n)$ is the singleton set $\{0\}$ which is not an event in $\mathcal{A}$. Accordingly, we have the following definition.

Definition 4. Given a set Ω, a *σ-algebra* (read "sigma-algebra") is an algebra of sets $\mathcal{F}$ that is also closed under countable unions: $A_i \in \mathcal{F}$ for $i \in \mathbb{N}$ implies that $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

The reader may have noticed that, as a consequence of the appropriate De Morgan's law, we immediately have the following proposition.

Proposition 5. *Let $\mathcal{F}$ be a σ-algebra. Then $\mathcal{F}$ is closed under countable intersections: $A_i \in \mathcal{F}$ for $i \in \mathbb{N}$ implies that $\bigcap_{i=1}^{\infty} A_i \in \mathcal{F}$.*

Given an algebra for a sample space Ω we can enlarge it so as to become a σ-algebra. In fact, any collection $\mathcal{S}$ of subsets of Ω admits a unique smallest σ-algebra, which we will denote $\sigma(\mathcal{S})$, that contains it. Since σ-algebras containing $\mathcal{S}$ exist (e.g., the power set of Ω) and arbitrary intersections of σ-algebras are again σ-algebras, we can define $\sigma(\mathcal{S})$ to be the intersection of all σ-algebras containing the set $\mathcal{S}$:

$$\sigma(\mathcal{S}) = \{A \subset \Omega \mid A \in \mathcal{F} \text{ for each } \sigma\text{-algebra } \mathcal{F} \text{ containing } \mathcal{S}\}.$$

Since $\sigma(\mathcal{S})$ is a σ-algebra, any σ-algebra that contains $\mathcal{S}$ must also contain $\sigma(\mathcal{S})$; in this sense, $\sigma(\mathcal{S})$ is smallest σ-algebra containing $\mathcal{S}$.

Definition 5 (Borel σ-algebra)**.** Let $\Omega = \mathbb{R}$ and let $\mathcal{S} = \{[a, b) \mid a, b \in \mathbb{R}\}$. The σ-algebra generated by $\mathcal{S}$ is called the Borel σ-algebra for $\mathbb{R}$ (or the σ-algebra of Borel sets for $\mathbb{R}$). The Borel σ-algebra for $\mathbb{R}$ is denoted by $\mathcal{B}(\mathbb{R})$, or often just $\mathcal{B}$.

The σ-algebra $\mathcal{B}(\mathbb{R})$ is large. For example, it contains all open intervals of the form (a, b). In fact, if $\mathcal{S}' = \{(a, b) \mid a, b \in \mathbb{R}\}$, then $\mathcal{B}(\mathbb{R})$ can be also be defined[6] as the σ-algebra generated by S'. Alternatively, we could use for the generating set S intervals of the form $(a, b]$, or intervals of the form $[a, b]$, or even intervals of the form $(-\infty, b]$, or $(-\infty, b)$, or $[a, \infty)$, or of the form (a, ∞); in each of those cases the resulting σ-algebra is $\mathcal{B}(\mathbb{R})$.

For those familiar with basic point set topology, $\mathcal{B}(\mathbb{R})$ is also generated by the open sets in $\mathbb{R}$ or, equivalently, by the closed sets (since all closed sets are complements of open sets). More generally, if X is any metric space then $\mathcal{B}(X)$, the Borel σ-algebra of X, is the σ-algebra generated by the open sets (or the closed sets) in X (and, if X is separable, $\mathcal{B}(X)$ is generated by the open balls in X).

Before leaving this section, let us prove the following technical result, which will be used below.

Lemma 1. *Let $f : \Omega \to \Omega'$ be a function (of sets) and suppose that Ω is equipped with a σ-algebra $\mathcal{F}$. Let $\mathcal{G} = \{B \subset \Omega' \mid f^{-1}(B) \in \mathcal{F}\}$. Then $\mathcal{G}$ is σ-algebra of sets in Ω'.*

Proof. First note that $f^{-1}(\Omega') = \Omega \in \mathcal{F}$, hence Ω' is in $\mathcal{G}$.

Let $B \in \mathcal{G}$. Then $f^{-1}(B) \in \mathcal{F}$ and, since $\mathcal{F}$ is a σ-algebra, we must have that $\Omega \setminus f^{-1}(B) \in \mathcal{F}$. But $f^{-1}(\Omega' \setminus B) = \Omega \setminus f^{-1}(B)$, hence $\Omega' \setminus B \in \mathcal{G}$ we see that $\mathcal{G}$ is closed under complementation.

Now suppose that $B_i \in \mathcal{G}$ for each $i \in \mathbb{N}$. Then for each i, $f^{-1}(B_i) \in \mathcal{F}$. But then $\bigcup_{i=1}^{\infty} f^{-1}(B_i) \in \mathcal{F}$ and, since $f^{-1}(\bigcup_{i=1}^{\infty} B_i) = \bigcup_{i=1}^{\infty} f^{-1}(B_i)$ we have that $\bigcup_{i=1}^{\infty} B_i \in \mathcal{F}$ and $\mathcal{G}$ is closed under countable unions. □

[6] To see that $\sigma(S') \subset \sigma(S)$ note that $\lim_{n\to\infty}[a + 1/n, b) = (a, b)$ and to see that $\sigma(S') \supset \sigma(S)$ notice that $\lim_{n\to\infty}(a - 1/n, b) = [a, b)$.

1.4 Probability spaces

Let Ω be a set and let $\mathcal{F}$ be σ-algebra of subsets of Ω. The pair $(\Omega,\, \mathcal{F})$ is called a *sample space* or a *measurable space*[7]. The sets in $\mathcal{F}$ are said to be *measurable*. The pair $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is a fundamental example of a measure space.

We can build new measure spaces from old by taking products. Let $(\Omega_1, \mathcal{F}_1)$, $(\Omega_2, \mathcal{F}_2)$, $\ldots$, $(\Omega_n, \mathcal{F}_n)$ each be measure spaces. We endow the product of sets $\Omega_1 \times \Omega_2 \times \cdots \times \Omega_n$ the structure of a measure space by defining its σ-algebra, which we denote by $\mathcal{F}_1 \times \mathcal{F}_2 \times \cdots \times \mathcal{F}_n$, to be one generated by all subset of the form $A_1 \times A_2 \times \cdots \times A_n$, where $A_i \in \mathcal{F}_i$ for all i. We define the product measure space by

$$(\Omega_1, \mathcal{F}_1) \times (\Omega_2, \mathcal{F}_2) \times \cdots \times (\Omega_n, \mathcal{F}_n) = (\Omega_1 \times \Omega_2 \times \cdots \times \Omega_n, \mathcal{F}_1 \times \mathcal{F}_2 \times \cdots \times \mathcal{F}_n).$$

Example 2. By the above construction $\mathbb{R}^n = \mathbb{R} \times \mathbb{R} \times \cdots \times \mathbb{R}$ can be endowed the structure of a measure space, denoted $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, by viewing it as the product of n copies of $(\mathbb{R}, \mathcal{B}(R))$. Further, the resulting product σ-algebra $\mathcal{B}(\mathbb{R}^n)$ is the same as that generated by the open balls in $\mathbb{R}^n$ (or by the closed balls) or by any basis for the usual topology on $\mathbb{R}^n$ (i.e., the metric topology for the Euclidean metric); it is also the same, for instance, as the σ-algebra generated by the sets of the form $(-\infty, b_1) \times (-\infty, b_2) \times \cdots \times (-\infty, b_n))$.

More generally, if $X_1, X_2, \ldots, X_n$ are separable metric spaces, then the product measure space $\mathcal{B}(X_1) \times \mathcal{B}(X_2) \times \cdots \times \mathcal{B}(X_n)$ coincides with the Borel σ-algebra $\mathcal{B}(X_1 \times X_2 \times \cdots \times X_n)$ generated by the open (or closed) sets in the product metric (which metrizes the product topology on $X_1 \times X_2 \times \cdots \times X_n$).

Now let $\mathcal{A}$ be an algebra of subsets of a sample space Ω. Recall that a probability measure is, by Definition 3, a map $P : \mathcal{A} \to \mathbb{R}$ that satisfies:

- $P(A) \geq 0$ for all $A \in \mathcal{A}$, and
- $\sum_{i=1}^{\infty} P(A_i) = 1$ whenever $\{A_i\}_{i=1}^{\infty}$ partitions Ω; i.e., whenever $A_i \in \mathcal{A}$ for all i, $\bigcup_{i=1}^{\infty} A_i = \Omega$, and the A_i are mutually disjoint.

The probability measure P extends uniquely to a map $P^* : \sigma(\mathcal{A}) \to \mathbb{R}$; meaning that P^* satisfies

(i) $P^*(A) \geq 0$ for all $A \in \sigma(\mathcal{A})$,

(ii) $\sum_{i=1}^{\infty} P^*(A_i) = 1$ whenever $\{A_i\}_{i=1}^{\infty}$ partitions Ω; i.e., whenever $A_i \in \sigma(\mathcal{A})$ for all i, $\bigcup_{i=1}^{\infty} A_i = \Omega$, and the $A_i \in \sigma(\mathcal{A})$ are mutually disjoint, and

(iii) $P^*(A) = P(A)$ when $A \in \mathcal{A}$.

Concretely, $P^* : \sigma(\mathcal{A}) \to \mathbb{R}$ is defined as follows. Let $A \in \sigma(\mathcal{A})$. Then $P^*(A) = \inf \sum_{i=1}^{\infty} P(A_i)$ where the infimum is over all families $\{A_i\}_{i=1}^{\infty}$ of subsets of Ω satisfying $A_i \in \mathcal{A}$ and $A \subset \bigcup_{i=1}^{\infty} A_i$.

Let us record this standard result (from measure theory) in the form a theorem.

[7] The word *measurable* references *measure theory*, the area of mathematics concerned with computing lengths in $\mathbb{R}$, areas in $\mathbb{R}^2$, volumes in $\mathbb{R}^3$, and so on. Intuitively, the *measure* of a set is its size (or, more accurately, mass, according to some mass distribution) and measure theory builds the foundations of such a theory at the ideal level of generality. In Calculus we use Riemann integration to compute areas, volumes, etc. Measure theory offers a more robust method — namely, *Lebesque integration* — for computing such (hyper-)volumes.

Theorem 1 (Carathéodory)**.** *A probability measure P defined on an algebra $\mathcal{A}$ has a unique extension to a probability measure P^* to $\sigma(\mathcal{A})$.*

We can now start working with probability spaces. First, compare the following definition with Definition 3 and Proposition 2, part (v).

Definition 6. Let $(\Omega, \mathcal{F})$ be a measureable space. A function $\mu : \Omega \to [0, \infty)$ is a *measure* if it is *countable additive*; that is, if $\mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$ whenever $A_i \in \mathcal{F}$ for all i and the A_i are mutually disjoint.

In measure theory, the triple $(\Omega, \mathcal{F}, \mu)$ is often referred to as a *measure space* — not to be confused with a *measureable* space.

Definition 3 defines a probability measure on algebras. But in order that monotonic sequences of measureable events converge to measureable events, we know that we should work on σ-algebras.

Definition 7. A measure P on a measureable space $(\Omega, \mathcal{F})$ is a *probability measure* if $P(\Omega) = 1$.

We now have, finally, Definition 1, which states that a probability space is a triple $(\Omega, \mathcal{F}, P)$ where P is a probability measure for the measureable space $(\Omega, \mathcal{F})$. Said differently, a probability space is a measure space with total measure $P(\Omega) = 1$. If $A \in \mathcal{F}$, then $P(A)$ is the *probability* of A.

1.5 Random variables

As usual in mathematics, we don't study just single objects — which, for now, are measureable spaces, and below, probability spaces. Rather, we study pairs of objects along with maps between them that preserve the relevant structure.

So, when we define functions mapping between sample (i.e., measurable) spaces, we are interested not in all possible (set-theoretic) functions but, rather, only such maps that *respect* the additional σ-algebra structure.

Definition 8. Let $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{F}')$ be measureable spaces. A function $f : \Omega \to \Omega'$ is *measureable* if $A \in \mathcal{F}'$ implies that $f^{-1}(A) \in \mathcal{F}$.

Here f^{-1} refers to the set-theoretic pre-image: $f^{-1}(A) = \{\omega \in \Omega \mid f(\omega) \in A\}$.

Proposition 6. *Let $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{F}')$ be measureable spaces, and let $\mathcal{S}$ be a family of subsets of Ω' that generates $\mathcal{F}'$; i.e., $\mathcal{F}' = \sigma(\mathcal{S})$. Then a function (of sets) $f : \Omega \to \Omega'$ is measureable if and only if $f^{-1}(A) \in \mathcal{F}$ for all $A \in \mathcal{S}$.*

Proof. If f is measureable, then $f^{-1}(A) \in \mathcal{F}$ for all $A \in \mathcal{F}$, and hence for all $A \in \mathcal{S}$ since $\mathcal{S} \subset \mathcal{F}$.

For the converse, suppose that $f^{-1}(A) \in \mathcal{F}$ for all $A \in \mathcal{S}$ and consider the set $\mathcal{G} \coloneqq \{B \subset \Omega' \mid f^{-1}(B) \in \mathcal{F}\}$. By Lemma 1, $\mathcal{G}$ is actually a σ-algebra for Ω' — and it contains $\mathcal{S}$. But $\mathcal{F}'$ is, by definition, the smallest σ-algebra containing $\mathcal{S}$. Hence $\mathcal{F}'$ is contained in $\mathcal{G}$. By definition of $\mathcal{G}$, we have, then, that $A \in \mathcal{F}' \Rightarrow f^{-1}(A) \in \mathcal{F}$, so that f is measureable. □

Definition 9 (Random Variable)**.** Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $(\Omega', \mathcal{F}')$ be a measureable space. A measureable function from Ω to Ω' is a called a *random variable.*

The codomain of random variables of interest is often $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ or $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$.

Lemma 2. *Let $f : \Omega \to \Omega'$ be measureable with respect to the $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{F}')$. Then the family of subsets $\mathcal{G} := \{f^{-1}(A) \mid A \in \mathcal{F}'\} \subset \mathcal{F}$ is a σ-algebra of subsets of Ω.*

Corollary 1. *Let $(\Omega, \mathcal{F}, P)$ be a probability space, and equip $\mathbb{R}$ with the Borel algebra $\mathcal{B}(\mathbb{R})$. Then a function $\mathcal{X} : \Omega \to \mathbb{R}$ is a random variable if and only if $\mathcal{X}^{-1}((-\infty, b)) \in \mathcal{F}$ for all $b \in \mathbb{R}$.*

More generally, $\mathcal{X} : \Omega \to \mathbb{R}^n$ is a random variable if and only if $\mathcal{X}^{-1}((-\infty, b_1) \times (-\infty, b_2) \times \cdots \times (-\infty, b_n)) \in \mathcal{F}$ for all $b_1, \ldots, b_n \in \mathbb{R}$.

Proof. This follows immediately from Proposition 6 since intervals of the form $(-\infty, b_1) \times (-\infty, b_2) \times \cdots \times (-\infty, b_n)$ generate the Borel algebra for $\mathbb{R}^n$. □

When working with random variables, we might suppress explicit mention of the probability measure and the σ-algebras involved. So if we write say $\mathcal{X} : \Omega \to \Omega'$ and specify that $\mathcal{X}$ is a random variable, we mean that there are σ-algebras $\mathcal{F}$ and $\mathcal{F}'$ so that $\mathcal{X} : \Omega \to \Omega'$ is measureable with respect to $\mathcal{F}$ and $\mathcal{F}'$; and that, in fact, $(\Omega, \mathcal{F})$ is equipped with the probability measure P that makes $(\Omega, \mathcal{F}, P)$ a probability space.

The reader may have noticed the asymmetry between the structure of the domain and codomain in the definition of a random variable: the domain is probability space while the codomain is a measureable space. There is no assumption of a probability measure on the codomain. However, a random variable $\mathcal{X} : \Omega \to \Omega'$ induces a probability measure on the codomain as follows.

Let $\mathcal{X} : \Omega \to \Omega'$ be a random variable with respect to the probability space $(\Omega, \mathcal{F}, P)$ and the measure space $(\Omega', \mathcal{F}')$. Then $\mathcal{X}$ canonically *induces* a probability measure, denoted $\mathcal{X}_{\#}P$, on its codomain: $\mathcal{X}_{\#}P : \Omega' \to [0, 1]$. The induced probability measure[8] is defined by $(\mathcal{X}_{\#}P)(A) = P(\mathcal{X}^{-1}(A))$ where $A \in \mathcal{F}'$.

Definition 10. Let $(\Omega, \mathcal{F}, P)$ and $(\Omega', \mathcal{F}', P')$ be probability spaces. A measureable map $f : \Omega \to \Omega'$ (of the underlying measure spaces $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{F}')$) is *measure-preserving* if $P(f^{-1}(A)) = P'(A)$ for all $A \in \mathcal{F}'$.

For example, given a random variable $\mathcal{X} : \Omega \to \Omega'$, the resulting map between probability spaces $(\Omega, \mathcal{F}, P)$ and $(\Omega', \mathcal{F}', P')$ is measure-preserving where P' is the induced measure $\mathcal{X}_{\#}P$.

1.6 Distributions

Definition 11. Let $\mathcal{X} : \Omega \to \mathbb{R}$ be a random variable. The *(cummulative) distribution function* (or CDF) $F_{\mathcal{X}} : \mathbb{R} \to [0, \infty)$ of $\mathcal{X}$ is given by $F_{\mathcal{X}}(x) = P(\mathcal{X} \leq x)$.

The notation $P(\mathcal{X} \leq x)$ is shorthand for $P(\mathcal{X}^{-1}(-\infty, x])$ where $\mathcal{X} : \Omega \to \mathbb{R}$ is a real-valued random variable (mapping between $(\Omega, \mathcal{F}, P)$ and $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$).

Proposition 7. *The distribution function $F_{\mathcal{X}}$ for a random variable $\mathcal{X} : \Omega \to \mathbb{R}$*

(i) is non-decreasing: $x_1 \leq x_2 \Rightarrow F_{\mathcal{X}}(x_1) \leq F_{\mathcal{X}}(x_1)$,

(ii) satisfies $\lim_{x \to -\infty} F_{\mathcal{X}}(x) = 0$ and $\lim_{x \to \infty} F_{\mathcal{X}}(x) = 1$, and

[8] Since $\mathcal{X}_{\#}P(\Omega') = P(\mathcal{X}^{-1}(\Omega')) = P(\Omega) = 1$, in order to check that $\mathcal{X}_{\#}P$ is in fact a probability measure we need only check that it is countable additive: if $\{A_i\}$ is a countable infinite collection of pairwise disjoint sets in $\mathcal{F}'$, then $\mathcal{X}_{\#}P(\bigcup_{i=1}^{\infty} A_i) = P(\mathcal{X}^{-1}(\bigcup_{i=1}^{\infty} A_i)) = P(\bigcup_{i=1}^{\infty} \mathcal{X}^{-1}(A_i)) = \bigcup_{i=1}^{\infty} P(\mathcal{X}^{-1}(A_i)) = \bigcup_{i=1}^{\infty} \mathcal{X}_{\#}P(A_i)$.

(iii) is right continuous: $\lim_{x\to x_0^+} F_{\mathcal{X}}(x) = F_{\mathcal{X}}(x_0)$ *for all* $x_0 \in \mathbb{R}$.

Proof. To prove part (iii), recall that a real-valued function f of a real variable is right continuous if any non-increasing sequence $\{x_n\}_{n=1}^{\infty}$ that converges to x_0 satisfies $\lim_{n\to\infty} f(x_n) = f(x_0)$. For such a sequence $\{x_n\}$ the sets $\mathcal{X}^{-1}((-\infty, x_n])$ for $n = 0, 1, \ldots$ are non-increasing and converge to $\mathcal{X}^{-1}((-\infty, x_0])$; hence $\lim_{n\to\infty} F_{\mathcal{X}}(x_n) = P(\mathcal{X} \leq x_n)$ □

1.7 (Optional) abstract nonsense

For the reader who happens to be familiar with basic category theory, the induced measure restores symmetry in the following sense. We have a category **Prob** whose objects are probability spaces and whose morphisms are almost-everywhere-equality measure preserving-maps. A measureable function $f : \Omega \to \Omega'$ mapping between probability spaces $(\Omega, \mathcal{F}, P)$ and $(\Omega', \mathcal{F}', P')$ is such a morphism[9] if $P'(A) = P(f^{-1}(A))$ for all measureable $A \subset \Omega'$. Notice that a random variable $\mathcal{X} : \Omega \to \Omega'$ is, then, a morphism in **Prob** if the measure on the codomain is the induced measure. **Prob**, however, is not the category in which probability theorists work.

Consider the set of all random variables $\mathcal{X} : \Omega \to \Psi$ that take values in a fixed measureable codomain $(\Psi, \mathcal{G})$. This set naturally forms a category, denoted $\mathbf{R}(\Psi)$, whose morphisms are measure-preserving maps: if $\mathcal{X}_1 : \Omega_1 \to \Psi$ is a random variable with domain $(\Omega_1, \mathcal{F}_1, P_1)$ and $\mathcal{X}_2 : \Omega_2 \to \Psi$ is a random variable with domain $(\Omega_2, \mathcal{F}_2, P_2)$, then a morphism from $\mathcal{X}_1$ to $\mathcal{X}_2$ is an — a priori measure-preserving (see below) — map $f : \Omega_1 \to \Omega_2$ that makes the following diagram commute (i.e., that satisfies $f \circ \mathcal{X}_2 \to \mathcal{X}_1$).

$$\begin{tikzcd}
\Omega_1 \arrow[dd, "f"'] \arrow[dr, "\mathcal{X}_1"] & \\
& \Psi \\
\Omega_2 \arrow[ur, "\mathcal{X}_2"'] &
\end{tikzcd}$$

Notice that if f is a measureable function that makes the above diagram commute, then then diagram

$$\begin{tikzcd}
f^{-1}(A) & \mathcal{F}_1 \arrow[dr, "P_1"] & \\
& & \mathbb{R} \\
A \arrow[uu, mapsto] & \mathcal{F}_2 \arrow[uu] \arrow[ur, "P_2"'] &
\end{tikzcd}$$

automatically commutes — i.e., f is measure-preserving — since $(\mathcal{X}_1)_{\#}(P_1) = (\mathcal{X}_2)_{\#}(P_2)$. Said differently, $\mathbf{R}$ is a slice category

2 Random variables in practice

Change X to rvX below here, for r.v.s

Example 3. Imagine flipping a possibly biased coin countably infinitely many times and recording the results. The sample space Ω is the set of all infinite strings of Hs and Ts, such as

$$HTTTHTHTTHHH\ldots.$$

[9] put a note here about almost everywhere equality

Let X be the random variable that assigns to a string in Ω the number of flips required to get *heads*. (Let us remove the string of all Ts from Ω.) For instance, $TTHTH\ldots \longmapsto 3$ under X.

The random variable X in Example 3 is *discrete* since it takes on only countably many values (in $\mathbb{R}$) — namely, the positive integers.

Let $p > 0$ be the probability of *heads* in a single flip of the coin in Example 3. Assuming that the flips are independent (no flip influences any other), the CDF for the random variable X in Example 3 is $F_X(x) = 1 - (1-p)^{\lfloor x \rfloor}$ if $x \geq 1$ and zero otherwise (here $\lfloor \cdot \rfloor$ denotes the floor function: $\lfloor x \rfloor$ is the greatest integer less than or equal to x). If, for instance, $x = 4$ then $F_X(x) = F_X(4)$ is the likelihood that by the fourth flip at least one *heads* has occurred. The complementary event is that no *heads* occurred in the first four flips; hence $F_X(4) = 1 - (1-p)^4$; similarly, $F_X(4.2) = 1 - (1-p)^4$.

Definition 12. For a discrete random variable X we define the *probability (mass) function* (or PDF) $f_X : \mathbb{R} \to [0,1]$ for X to be $f_X(x) = P(X = x)$.

The CDF and PDF of a discrete random variable are related by

$$F_X(x) = \sum_{x_i \leq x} f_X(x_i).$$

In Example 3, the probability of getting *tails* on the first $n-1$ flips is $(1-p)^{n-1}$ while the probability of getting *heads* on the nth flip is p. Hence, $f_X(n) = p(1-p)^{n-1}$ where n is a positive integer.

Given a discrete random variable X, let $\{x_1, x_2, \ldots\}$ denote the complete list of real numbers taken on by X. Then X defines a probability space with sample space $\{x_1, x_2, \ldots\}$. The fact that the induced probability measure is σ-additive follows from basic facts about functions (for example, the pre-images of disjoint sets are disjoint). This is what is meant by *push forward*. A discrete random variable $X : \Omega \to \mathbb{R}$ allows us to push forward a probability measure P on Ω to a probability measure on the image of X in $\mathbb{R}$.

Let us use the random variable in Example 3 to test drive the claim that the push forward is an actual probability measure by checking that we, at least, have that $\sum_{n=1}^{\infty} P(x = n) = \sum_{n=1}^{\infty} f_X(n)$ is equal to 1:

$$\begin{aligned}\sum_{n=1}^{\infty} f_X(n) &= \sum_{n=1}^{\infty} p(1-p)^{n-1} \\ &= p\sum_{n=1}^{\infty} (1-p)^{n-1} \\ &= p\frac{1}{1-(1-p)} = 1\end{aligned}$$

where we have used the formula $\sum_{n=1}^{\infty} r^{n-1} = 1/(1-r)$ for summing a convergent (i.e., when $|r| < 1$) geometric series. Note that $0 < p \leq 1$ implies that $|1-p| < 1$.

We are often interested in some specific aspect of an experiment — such as the first occurrence of *heads* in a sequence of coin flips. As in Example 3, we can use an appropriate random variable to focus on the relevant phenomenon, thereby suppressing reference to the domain Ω of our random variable. Note however that, once X is chosen, the underlying probability space and probability measure ultimately determine a probability measure on $\mathbb{R}$ specified by X.

2.1 Discrete random variables

Definition 13. The *expected value* of a discrete random variable X is

$$E(X) = \sum_i x_i f_X(x_i) = \sum_i P(X = x_i).$$

For the random variable in Example 3, we have

$$\begin{aligned} E(X) = \sum_{n=1}^{\infty} n f_X(n) &= \sum_{n=1}^{\infty} np(1-p)^{n-1} \\ &= p \sum_{n=1}^{\infty} n(1-p)^{n-1} \\ &= p \frac{1}{(1-(1-p))^2} = \frac{1}{p}, \end{aligned}$$

where we have used the identity

$$\sum_{n=1}^{\infty} nx^{n-1} = \frac{d}{dx}\left(\sum_{n=1}^{\infty} x^n\right) = \frac{d}{dx}\left(x \sum_{n=1}^{\infty} x^{n-1}\right) = \frac{d}{dx}\left(\frac{x}{1-x}\right) = \frac{1}{(1-x)^2} \tag{4}$$

if $|x| < 1$, which follows from that fact that we can differentiate power series term-wise where they converge.

For instance, if $p = 0.4$ then over many trials (of continually flipping the coin), the first *heads* occurs, an average, after $E(0.4) = 2.5$ flips.

We can do arithmetic with random variables that are defined on the same probability space: if X and Y both map $\Omega \to \mathbb{R}$, then XY, $X+Y$, $3X$, $\frac{1}{2}Y$, etc. each define a new random variable $\Omega \to \mathbb{R}$ by performing the indicated operation in the range.

Expectation, acting as an operator on random variables, is linear: $E(aX + bY) = aE(X) + bE(Y)$ for $a, b \in \mathbb{R}$.

Definition 14. The *variance* of a random variable X is

$$V(X) = E\left((X - E(X))^2\right).$$

In general, since $E(X)$ is constant, we can write

$$\begin{aligned} V(X) &= E\left((X - E(X))(X - E(X))\right) \\ &= E\left(X^2 - 2E(X)X + (E(X))^2\right) \\ &= E\left(X^2\right) - E(2E(X)X) + E((E(X))^2) \\ &= E\left(X^2\right) - 2E(X)E(X) + (E(X))^2 \\ &= E\left(X^2\right) - E(X)^2. \end{aligned}$$

For the random variable in Example 3, we can use power series manipulations as follows to directly compute the variance.

First we compute the sum, to be used below, of the power series $\sum_{n=1}^{\infty} n^2 x^{n-1}$, which converges if $|x| < 1$. We start with

$$\frac{x^2}{1-x} = x^2 \frac{1}{1-x} = x^2 \sum_{n=1}^{\infty} x^{n-1} = \sum_{n=1}^{\infty} x^{n+1}. \tag{5}$$

The second derivative of the left side of (5) is $2/(1-x)^3$. Twice differentiating the right side of (5) term-wise, we arrive at $\sum_{n=1}^{\infty} n(n+1)x^{n-1} = \sum_{n=1}^{\infty} n^2x^{n-1} + \sum_{n=1}^{\infty} nx^{n-1}$ if $|x| < 1$. Hence, if $|x| < 1$,

$$\begin{aligned}\sum_{n=1}^{\infty} n^2x^{n-1} &= \frac{2}{(1-x)^3} - \sum_{n=1}^{\infty} nx^{n-1} \\ &= \frac{2}{(1-x)^3} - \frac{1}{(1-x)^2} \\ &= \frac{1+x}{(1-x)^3} \qquad (6)\end{aligned}$$

where we have used (4).

Next we compute, for X as in Example 3,

$$\begin{aligned}E\left(X^2\right) &= \sum_{n=1}^{\infty} n^2p(1-p)^{n-1} \\ &= p\sum_{n=1}^{\infty} n^2(1-p)^{n-1} \\ &= p\frac{1+(1-p)}{(1-(1-p))^3} = \frac{2-p}{p^2}\end{aligned}$$

where we used (6).

Hence, for our random variable X in Example 3,

$$\begin{aligned}V(X) &= E\left(X^2\right) - (E(X))^2 \\ &= \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}\end{aligned}$$

If, for instance, $p = 0.4$ then the expected number of flips to achieve *heads* is $1/0.4 = 2.5$, with variance $(1-0.4)/(0.4)^2 = 3.75$.

The expected value of a random variable is its one number summary. It is the average value of the random variable over many trials. From its definition, we see that the variance of random variable is a measure of the spread of the random variable's values around its expected value.

Later we will quantify the notion of *spread* in the context of *inference*. For now notice that, in Example 3, the variance approaches ∞ as the probability p of heads approaches 0 — corresponding, perhaps, with intuition.

The random variable occurring in Example 3 arises in experiments beyond coin flipping. If the essentially the same random variable occurs in different contexts (say from pushing forward from nominally different sample spaces), it makes sense to single that variable out mathematically.

We say that two random variables $X : \Omega_1 \to \mathbb{R}$ and $Y : \Omega_2 \to \mathbb{R}$ are *equal in distribution* if their CDFs are equal — that is, if $F_X(x) = F_Y(x)$ for all $x \in \mathbb{R}$; in this case, we write $X \sim Y$.

Definitions (important discrete probability distributions)**.** Let n be a non-negative integer and let p a non-negative real number.

Bernoulli random variable Bern(p)

binomial binomial(n,p)

The *geometric distribution*, geom(p)

Poisson(λ)

Now suppose that we have two random variables $X : \Omega_1 \to \mathbb{R}$ and $Y : \Omega_2 \to \mathbb{R}$, where Ω_1 and Ω_2 might be from completely different probability spaces.

Definition 15. Let X and Y be discrete random variables. The *joint mass function*, denoted $f_{X,Y}$ is the function mapping $\mathbb{R} \times \mathbb{R} \to [0,1]$ defined by $f_{X,Y}(x,y) = P(X = x, Y = y)$. Here, $P(X = x, Y = y) = P(X = x \text{ and } Y = y)$ denotes the probability that $X = x$ and $Y = y$.

Here we can think of the pair (X,Y) as a (generalized) random variable $\Omega_1 \times \Omega_2 \to \mathbb{R} \times \mathbb{R}$. Note the image of this map is discrete.

Example 4. Suppose that you continually flip a biased quarter and that each time you flip it your friend flips a biased nickel that he has. If your quarter comes up *heads* with probability p_1 and your friend's nickel comes up *heads* with probability p_2. What is the likelihood that both coins first come up heads on the same flip?

Let $X \sim \text{geom}(p_1)$ and $Y \sim \text{geom}(p_2)$ and consider the pair (X, Y). Then the probability we want is $\sum_{n=1}^{\infty} f_{X,Y}(n,n)$. Now, since an outcome for one coin has no bearing on the outcome of the other, $P(X = x, Y = y) = P(X = x)P(Y = y)$. Hence

$$\begin{aligned}
\sum_{n=1}^{\infty} f_{X,Y}(n,n) &= \sum_{n=1}^{\infty} p_1(1-p_1)^{n-1} p_2(1-p_2)^{n-1} \\
&= p_1 p_2 \sum_{n=1}^{\infty} ((1-p_1)(1-p_2))^{n-1} \\
&= \frac{p_1 p_2}{1-(1-p_1)(1-p_2)} \\
&= \frac{p_1 p_2}{p_1 + p_2 - p_1 p_2}
\end{aligned}$$

For instance, if $p_1 = p_2 = 0.5$ (two fair coins) the probably that they first come up heads on the same toss is $1/3$.

Definition 16. Two random variables X and Y are *independent* if $P(X = x, Y = y) = P(X = x)P(Y = y)$.

Proposition 8. *Two random variables X and Y are independent if and only if their joint mass function satisfies $f_{X,Y}(x,y) = f_X(x) f_Y(y)$.*

Proof. If X and Y be independent, then $f_{X,Y}(x,y) = P(X = x, Y = y) = P(X = x)P(Y = y) = f_X(x) f_Y(y)$. Conversely, if $f_{X,Y}(x,y) = f_X(x) f_Y(y)$, then $P(X = x, Y = y) = f_{X,Y}(x,y) = f_X(x) f_Y(y) = P(X = x)P(Y = y)$. □

The random variables in the last example are independent. In fact, we can take any two of the random variables defined above, declare that they are independent, and compute their joint mass function via $f_{X,Y}(x,y) = f_X(x) f_Y(y)$.

Example 5. Suppose now that, flipping a quarter and a nickel as in Example 4, we are interested in the flip that, on average, first produces *heads* on the nickel when the corresponding flip of the quarter is also (not necessarily the first) *heads*.

Definition 17. Given a pair $(X, Y) : \Omega_1 \times \Omega_2 \to \mathbb{R} \times \mathbb{R}$ of random variables $X : \Omega_1 \to \mathbb{R}$ and $Y : \Omega_2 \to \mathbb{R}$, the generalized random variable X conditioned on Y, denoted $X|Y$, is defined to be the map $X|Y : \Omega_1 \times \Omega_2 \to \mathbb{R} \times \mathbb{R}$ with probability mass function $f_{X|Y}(x, y) = f_{X,Y}(x, y)/f_Y(y)$.

Returning to Example 5, let $X \sim \text{Bern}(p_1)$ and $Y \sim \text{geom}(p_2)$. Then X represents a single flip of the quarter and Y is flip for the first *heads* while continually flipping the nickel. We are interested in $E(Y|X = 1)$.

$$E(Y|X = 1) = \sum_{n=1}^{\infty} f_{Y|X}()$$

$$\vdots$$

A random variable X is *continuous* if it admits a

So a random variable X is a map $X : \Omega \to \mathbb{R}$, and a probability measure P is a map $P : \mathcal{F} \to [0, \infty)$.

$$f_X$$

References

[1] https://www.encyclopediaofmath.org/index.php/Algebra_of_sets

[2] https://en.wikipedia.org/wiki/Measure_(mathematics)

[3] https://en.wikipedia.org/wiki/Sigma-algebra

[4] https://www.encyclopediaofmath.org/index.php/Measure

[5] https://en.wikipedia.org/wiki/Set-theoretic_limit

[6] S.R.S. Varadhan, *Probability Theory*, Courant Lecture Notes 7.

[7] golem.ph.utexas.edu...a_categorical_look_at_random_v.

[8] mathoverflow.net...structuralist-categorical