Probability 1 lecture notes

Clare Wallace

2025-10-02

Table of contents

Welcome to Probability 1

Welcome to Probability 1! These lecture notes contain all the mathematical content you'll need to know to succeed in Probability this year.

If you have questions about any of the content here, try one of the following:

- ask a friend!
- ask me! I like to answer emails, and I am often in my office (MCS3060): you can (and should) pop by to see if I'm around. You can do this during my official office hours (Mondays, 10-12) for a guaranteed speedy response, but you definitely shouldn't wait until then, especially if it's a short or quick question.
- Google it, or try a textbook. There are some good ones on the reading list (see below).

These notes have been developed over the years by several members of the Statistics and Probability groups, including (most recently) Debleena Thacker and Andrew Wade.

Warning

There could still be typos. If you find one, let me know about it and you can have a free bag of Skittles.

How to use these notes

The notes contain all the mathematical content for the course. In lectures, we will start at the beginning and work our way through the whole document, until we reach the end (hopefully, this will happen exactly at the end of term).

Throughout the notes, there are boxes like this one:

Try it out

You can do the "Introductory" exercises on the problem sheet already.

These contain examples you can work through to check your understanding. Wherever possible, I've also worked examples into the text, but there are some places where I want to give you an extra example. These come in purple boxes.

Content that's particularly important for the course is highlighted in red:

Key idea

Probability is cooler than statistics

while advanced material is highlighted in blue:

Advanced content

Probability is *almost surely* cooler than statistics.

You'll also find textbook recommendations, with the relevant sections:

Textbook references

If you want more help with this section, check out:

- Appendix A.1 in (Blitzstein and Hwang 2019);
- Appendix B in (Anderson, Seppäläinen, and Valkó 2018);
- or the Appendix to Chapter 0 in (Stirzaker 2003).

The library has lots of good books on introductory probability, and there are even more available online/to buy. The following four textbooks are a good starting point:

- (Blitzstein and Hwang 2019) covers the material in depth and uses simulation code to illustrate the theory.
- (Anderson, Seppäläinen, and Valkó 2018) covers just about everything in the course at about the right level of detail.
- (Stirzaker 2003) is concise and the most mathematically advanced, and will be useful for students taking 2H probability.
- (DeGroot and Schervish 2013) has a statistical perspective, covering this course as well as a lot of Statistics.

1 Axioms of probability

Goals

1. Understand elementary set theory and how to use it to formulate probabilistic scenarios and to describe the calculus of events.

2. Be familiar with the axioms of probability and their consequences, and how these properties may be deduced from the axioms.

In this chapter, we lay the foundations of probability calculus, and establish the main techniques for practical calculations with probabilities. The mathematical theory of probability is based on *axioms*, like Euclidean geometry. In classical geometry, the fundamental objects posited by the axioms are points and lines; in probability, they are *events* and their *probabilities.* The language and apparatus of set theory is used to express these concepts and to work with them.

There is a lot of ambiguity inherent in probability, because we are often using mathematical approaches to describe real-world scenarios. In some cases, there are several different ways to represent the real-world scenario as a probabilistic model, and the choices we make could affect our conclusions. In others, an unambiguous mathematical setup could have different real-world interpretations, depending on how we view it. Either way, once we have a probabilistic model, the axioms help us to ensure that the maths remains the same.

The axioms and properties of probability we develop in this chapter lay the foundations for all the rest of the theory we will build later in the course.

1.1 Sets

One of the key tools we need in this chapter is a good understanding of *set theory.* You'll see all of this much more formally in Analysis, but in this section we give a quick rundown of the essentials we need for Probability.

In essence, a set is an unordered collection of distinguishable objects; these objects can be numbers, functions, other sets, and so on—any mathematical object can belong to a set.

The formal notation for a set is an opening curly bracket, followed by a list of *elements* that belong to the set, followed by a closing curly bracket. For instance, the set containing the elements 2, 4, and 5 is denoted by

$$\{2, 4, 5\}.$$

Because the ordering of the elements is irrelevant, $\{2, 4, 5\}$ and $\{4, 5, 2\}$ denote the same set.

Definition: empty set

The set with no outcomes is called the *empty set*, and is denoted by $\emptyset$:

$$\emptyset := \{\}.$$

A set is often denoted by a capital letter such as A, B, C, and so on.

Definition: subset

For two sets A and B, we say that A is a *subset* of B, and we write $A \subseteq B$ (or $B \supseteq A$), whenever every element that belongs to A also belongs to B, that is, for all $x \in A$ we have $x \in B$.

For instance, $\{2, 4, 5\} \subseteq \{1, 2, 3, 4, 5\}$. Note that for every set A, we have $A \subseteq A$ and $\emptyset \subseteq A$. We can also use *strict* subsets, when the subset is not equal to the larger set: $\{2, 4, 5\} \subset \{1, 2, 3, 4, 5\}$.

Definition: power set

The set consisting of all subsets of a set A is called the *power set* of A, and is denoted as 2^A:

$$2^A := \{B \colon B \subseteq A\}.$$

For example, the power set of the set $A = \{1, 2, 3\}$ is

$$2^A = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}.$$

The notation 2^A alludes to the size of the power set. When A is a finite set, its power set contains $2^{|A|}$ subsets. This can be proved by constructing a bijection from 2^A to ordered $|A|$-tuples of 0s and 1s, where a 1 indicates that the corresponding element of A is in the subset.

Textbook references

If you want more help with this section, check out:

- Appendix A.1 in (Blitzstein and Hwang 2019);
- Appendix B in (Anderson, Seppäläinen, and Valkó 2018);
- or the Appendix to Chapter 0 in (Stirzaker 2003).

1.2 Sample space and events

Definition: scenarios, outcomes and sample space

Whenever we do some probability, it is based on a *scenario* in which there are various *outcomes*. We assume that we know the (set of all) possible outcomes, but we are unsure about which outcome will occur.
A *sample space* is a set of outcomes for this scenario with the property that one (and only one) of these outcomes must occur.
In this course, we will usually denote the sample space by Ω, and a generic outcome by $\omega \in \Omega$.

For instance, suppose we roll a standard six-sided die.

The most obvious sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$, but if one was interested only in whether the die was odd or even, or a six or not, one could use $\Omega = \{\text{odd}, \text{even}\}$, or $\Omega = \{\text{not a } 6, 6\}$.

Often, like in the above example, we may enumerate the elements of the sample space Ω in a finite or infinite list $\Omega = \{\omega_1, \omega_2, \ldots\}$, in which case we say the set Ω is *countable* or *discrete*.

A set is said to be countable when its elements can be enumerated in a (possibly infinite) sequence. Every finite set is countable, and so is the set of natural numbers $\mathbb{N} := \{1, 2, 3, \ldots\}$. The set of integers $\mathbb{Z}$ is countable as well. The set of real numbers $\mathbb{R}$ is not countable, and neither is any interval $[a, b]$ when $a < b$.

Definition: countable

A set A is countable if either:

- A is finite, or
- there is a bijection (one-to-one and onto mapping) between A and the set of natural numbers $\mathbb{N}$.

One can prove that the set of rational numbers $\mathbb{Q}$ is countable.

When we perform an experiment we are interested in the occurence, or otherwise, of *events*. An *event* is just a collection of possible outcomes, i.e., a subset of Ω.

Key idea: Definition: events

Associated to our sample space Ω is a collection $\mathcal{F}$ of *events*:

$$A \subseteq \Omega \text{ for every } A \in \mathcal{F}.$$

We say that an event A *occurs* when the outcome that occurs at the end of the scenario is in the set A.

If Ω is discrete, we can always take $\mathcal{F} = 2^{\Omega}$, so that *every* subset of Ω is an event. If Ω is not discrete, we need to be a little more careful: see Section 1.4 below.

The empty set $\emptyset$ represents the *impossible event*, i.e., it will never occur. The sample space Ω represents the *certain event*, i.e., it will always occur. Most interesting events are somewhere in between.

The representation of an event as a set obviously depends on the choice of sample space Ω for the specific scenario under study, as shown by the following two examples.

Examples

1. For rolling a standard cubic die (with $\Omega = \{1, 2, 3, 4, 5, 6\}$), the event “throw an odd number” is the subset $A = \{1, 3, 5\}$ consisting of three outcomes. If we roll the die and it comes up a 3, then event A has occurred.

2. For the same scenario, but with $\Omega = \{\text{odd}, \text{even}\}$, the event ‘throw an odd number’ is the subset $A = \{\text{odd}\}$ consisting of just one outcome.

Try it out

Suppose we roll three standard six-sided dice and record the outcome of the experiment by an ordered triple (i, j, k) where $i, j, k \in \{1, 2, \dots, 6\}$. What is Ω? How big could $\mathcal{F}$ be?
Answer:
In this case $\Omega = \{1, 2, \dots, 6\}^3$ has $6^3 = 216$ individual outcomes.
The number of events (2^{216}) is enormous! One is the event that the scores on the three dice are the same: $A = \{(1,1,1), (2,2,2), \dots, (6,6,6)\}$.

Textbook references

If you want more help with this section, check out:

- Section 1.2 in (Blitzstein and Hwang 2019);
- Section 1.1 in (Anderson, Seppäläinen, and Valkó 2018);
- Sections 1.1 and 1.2 in (Stirzaker 2003);
- or Section 1.4 in (DeGroot and Schervish 2013).

1.3 Event calculus

Once we've defined our sample space and the set of all possible events, we need to be able to refer to combinations of events. To do so, we use standard notation from set theory.

Definition: complements

For an event $A \in \mathcal{F}$, we define its *complement*, denoted A^{c} (or sometimes $\bar{A}$) and read "not A", to be

$$A^{\mathrm{c}} := \Omega \backslash A = \{\omega \in \Omega : \omega \notin A\}.$$

Notice that:

- the complement of A^{c} is A: $(A^{\mathrm{c}})^{\mathrm{c}} = A$;
- there are no outcomes in *both* A and A^{c}: $A \cap A^{\mathrm{c}} = \emptyset$;
- and every outcome is in one or the other: $A \cup A^{\mathrm{c}} = \Omega$.

Key idea: event calculus

Given any two events A and B that are associated with the same sample space (i.e. $A \subseteq \Omega$ and $B \subseteq \Omega$ for the same Ω), here are some of the other events we can define, along with how we would read them out:

Notation	We say (as sets)	We say (as events)	Meaning (as events)
$A \cup B$	A union B	A or B	A occurs or B occurs or both A and B occur
$A \cap B$	A intersect B	A and B	A occurs and B occurs
$A^{\mathrm{c}} := \Omega \backslash A$	A complement	not A	A does not occur
$A \backslash B$	A minus B	A but not B	A occurs but B does not
$A \subseteq B$	A is a subset of B	A implies B	if A occurs, then B must occur

(In the final row, "$A \subseteq B$" is not an event but rather a statement about how two events relate to each other. I still wanted to include it because I think it's helpful)

Try it out

Prove that $A \backslash B = A \cap B^{\mathrm{c}}$.
Answer:
We can do this by working with the events as sets. We have

$$A \backslash B = \{\omega \in \Omega : \omega \in A,\, \omega \notin B\} = \{\omega \in \Omega : \omega \in A\} \cap \{\omega \in \Omega : \omega \notin B\} = A \cap B^{\mathrm{c}}.$$

Key idea: Definition: disjoint

We say that events A and B are *disjoint*, *mutually exclusive*, or *incompatible* if $A \cap B = \emptyset$, i.e., it is impossible for A and B both to occur.

Try it out

Consider the sample space $\Omega := \{1, 2, 3, 4, 5, 6\}$, and the events

$$\begin{aligned} A &:= \{2, 4, 6\}, \\ B &:= \{1, 3, 5\}, \\ C &:= \{1, 2, 3\}. \end{aligned}$$

In other words, A is the event "throw an even number", B is the event "throw an odd number", and C is the event "throw at most three". Use some of the ideas from the table above to combine events A, B, and C in different ways. Are any of your new events disjoint?
Answer:
Some combinations:

$$\begin{aligned} A \cup B &= \Omega && \text{(even or odd)} \\ A \cap B &= \emptyset && \text{(even and odd)} \\ A^{\mathrm{c}} &= B && \text{(not even)} \\ C \backslash A &= \{1, 3\} && \text{(at most 3 but not even)} \\ A \cup C &= \{1, 2, 3, 4, 6\} && \text{(even or at most 3)} \\ A \cap C &= \{2\} && \text{(even and at most 3).} \end{aligned}$$

The events A and B are disjoint as $A \cap B = \emptyset$. We have also created two disjoint events: $C \backslash A$ and $A \cap C$. Think about why these two events would always be disjoint, however we define A and C.

Try it out

Toss a coin twice and denote the sample space by $\Omega = \{\mathrm{HH}, \mathrm{HT}, \mathrm{TH}, \mathrm{TT}\}$. Consider the events

$$\begin{aligned} A &:= \{\mathrm{HH}, \mathrm{HT}\} && \text{(first toss H)} \\ B &:= \{\mathrm{HT}, \mathrm{TT}\} && \text{(second toss T)} \\ C &:= \{\mathrm{HH}\} && \text{(both H).} \end{aligned}$$

How do these events relate to each other?
Answer:
Some things you might notice:

- $C \subseteq A$, i.e., if C occurs then A must occur;
- $A \cup B = \{\mathrm{HH}, \mathrm{HT}, \mathrm{TT}\}$ is the event that either the first toss is H, the second toss is T, or both;
- $A \cap B = \{\mathrm{HT}\}$;
- $A^{\mathrm{c}} = \{\mathrm{TH}, \mathrm{TT}\}$;
- $B \cap C = \emptyset$.

Try it out

Draw a card from a standard deck of 52 playing cards. Take Ω to consist of each of the 52 possible draws: $\Omega = \{\mathrm{A}\clubsuit, \mathrm{A}\diamondsuit, \dots, \mathrm{K}\heartsuit, \mathrm{K}\spadesuit\}$. Events in $\mathcal{F} = 2^{\Omega}$ include

$$\begin{aligned} E &= \{\text{eight}\} = \{8\spadesuit, 8\heartsuit, 8\diamondsuit, 8\clubsuit\}, \\ S &= \{\text{spade}\} = \{A\spadesuit, 2\spadesuit, \dots, K\spadesuit\}, \end{aligned}$$

and we can combine them to form other events, such as

$$\begin{aligned} E \cap S &= \{8\spadesuit\}, \\ E\backslash S &= \{8\heartsuit, 8\diamondsuit, 8\clubsuit\}, \\ S\backslash E &= \{A\spadesuit, 2\spadesuit, \dots, 7\spadesuit, 9\spadesuit, \dots, K\spadesuit\}. \end{aligned}$$

As with sums ($\sum$) and products (Π) of multiple numbers, we also have shorthands for unions and intersections of multiple sets:

$$\bigcup_{i=1}^{n} A_i := A_1 \cup A_2 \cup \cdots \cup A_n$$

is the event that at least one of $A_1, A_2, \dots A_n$ occurs (or the set of all $\omega \in \Omega$ which are contained in at least one of the A_is), and

$$\bigcap_{i=1}^{n} A_i := A_1 \cap A_2 \cap \cdots \cap A_n$$

is the event that *all* of $A_1, A_2, \dots A_n$ occur (or the set of all $\omega \in \Omega$ which are in every A_i).

Occasionally, we will also need to take infinite unions and intersections over sequences of sets:

$$\begin{aligned} \bigcup_{i=1}^{\infty} A_i &:= A_1 \cup A_2 \cup A_3 \cup \dots \\ \bigcap_{i=1}^{\infty} A_i &:= A_1 \cap A_2 \cap A_3 \cap \dots . \end{aligned}$$

We will also sometimes need *De Morgan's Laws*: for a (possibly infinite) collection of events A_i,

a. The complement of the union is the intersection of the complements: $\left(\bigcup_i A_i\right)^c = \bigcap_i A_i^c$, and
b. The complement of the intersection is the union of the complements: $\left(\bigcap_i A_i\right)^c = \bigcup_i A_i^c$.

These could be more intuitive than they appear: the negation of "some of these things happened" is "none of these things happened", and the negation of "all of these things happened" is "some of these things did not happen".

It is often useful to visualize the sample space in a *Venn diagram*. Then events such as A are subsets of the sample space. It is a helpful analogy to imagine the probability of an event as the *area* in the Venn diagram.

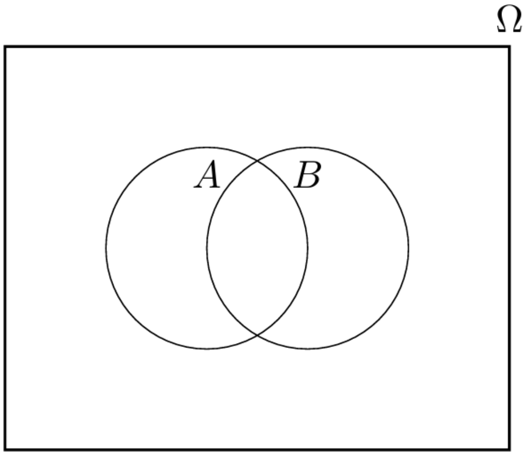


Figure 1.1: Venn diagram

Advanced content

This analogy is more apt than it first appears, since the mathematical foundations of rigorous probability theory are built on *measure theory*, which is the same theory that gives rigorous foundation to the concepts of length, area, and volume.

Textbook references

If you want more help with this section, check out:

- Section 1.2 in (Blitzstein and Hwang 2019);
- Section 1.2 in (Stirzaker 2003);
- or Section 1.4 in (DeGroot and Schervish 2013).

1.4 Sigma-algebras

In the last section we described some of the ways in which events can be combined. Now we can set out the rules for our collection of events, $\mathcal{F}$, to ensure that it's possible to *use* these different combinations.

We said that in the case where Ω is discrete, one can take $\mathcal{F} = 2^\Omega$.

In general, if Ω is uncountable, it is too much to demand that probabilities should be defined on *all* subsets of Ω. The reason why this is a problem goes beyond the scope of this course (see the Bibliographical notes at the end of this chapter for references), but the essence is that for uncountable sample spaces, such as $\Omega = [0, 1]$, there exist subsets of Ω that cannot be assigned a probability in a way that is consistent. The construction of such *non-measurable sets* is also the basis of the famous *Banach–Tarski paradox*.

Uncountable Ω are unavoidable: we will see an infinite coin-tossing space at the end of section Section 1.6, and other examples occur whenever we have an experiment whose outcome is modelled by a continuous distribution such as the normal distribution (more on this later).

The upshot of all this is that we can, in general, only demand that probabilities are defined for all events in some collection $\mathcal{F}$ of subsets of Ω (i.e., for some $\mathcal{F} \subseteq 2^{\Omega}$). What properties should the collection $\mathcal{F}$ of events possess? Consideration of the set operations in the previous section suggests the following definition.

Definition: -algebra

A collection $\mathcal{F}$ of subsets of Ω is called a *σ-algebra* over Ω if it satisfies the following properties.
(S1) $\Omega \in \mathcal{F}$;
(S2) $A \in \mathcal{F}$ implies that $A^{\mathrm{c}} \in \mathcal{F}$;
(S3) if $A_1, A_2, \dots \in \mathcal{F}$ then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Property **S2** says that $\mathcal{F}$ is *closed under complementation*, while **S3** says that $\mathcal{F}$ is *closed under countable unions*.

We can combine **S1** and **S2** to see that we must have $\emptyset \in \mathcal{F}$. Also note that, we can get to a finite-union version of **S3** by taking $A_{n+1} = A_{n+2} = \dots = \emptyset$: so $\mathcal{F}$ is also closed under finite unions.

Examples

1. The power set 2^{Ω} is a σ-algebra over Ω, and in fact it is the biggest possible σ-algebra over Ω. As described above, for uncountable Ω the set 2^{Ω} may be too unwieldy, in which case we would consider a smaller σ-algebra.

2. The *trivial σ-algebra* $\{\emptyset, \Omega\}$ is the smallest possible σ-algebra over Ω. It's very nicely behaved (just two elements!) but it carries no information about the outcome of the experiment.

Try it out

Consider the sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$ for the experiment of rolling a fair die. The choice of σ-algebra determines the *resolution at which we observe the experiment*, and may depend on exactly what we are interested in:

- $\mathcal{F}_0 = \{\emptyset, \Omega\}$ (carries no information);
- $\mathcal{F}_1 = \{\emptyset, \{1, 3, 5\}, \{2, 4, 6\}, \Omega\}$ (if we only care whether the score is odd or even);
- $\mathcal{F}_2 = 2^{\Omega}$ (if we are interested in the exact score).

Note the inclusions $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2$.
Let us check that $\mathcal{F}_1$ is indeed a σ-algebra.
S1 is immediate: we can see it from the definition of $\mathcal{F}_1$.
For **S2**, we need that every $A \in \mathcal{F}_1$ is accompanied by its complement A^{c}; we see that this is the case.

Since $\mathcal{F}_1$ is a finite set it suffices to check **S3** for finite unions. In other words, it is enough to check that if $A, B \in \mathcal{F}_1$, then $A \cup B \in \mathcal{F}_1$ too. Since there are only two sets, this is also quick: we see that it is the case.

Textbook references

If you want more help with this section, check out:

- Section 1.2 in (Stirzaker 2003).

1.5 The axioms of probability

Key idea: Definition: probability

A *probability* $\mathbb{P}$ on a sample space Ω with collection $\mathcal{F}$ of events is a function mapping every event $A \in \mathcal{F}$ to a real number $\mathbb{P}(A)$, obeying the following axioms:
(A1) $\mathbb{P}(A) \geq 0$ for every $A \in \mathcal{F}$;
(A2) $\mathbb{P}(\Omega) = 1$; and
(A3) if A and B are *disjoint* events (i.e. if $A, B \in \mathcal{F}$ have $A \cap B = \emptyset$) then

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B).$$

We call the number $\mathbb{P}(A)$ *the probability of* A.

We will see shortly that a consequence of these axioms is that the probabilities $\mathbb{P}(A)$ must lie between 0 and 1: $0 \leq \mathbb{P}(A) \leq 1$.

We can upgrade **(A3)** to a slightly more technical version:

(A4) For any infinite sequence $A_1, A_2, \ldots$ of pairwise disjoint events (so $A_i \cap A_j = \emptyset$ for all $i \neq j$),

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

Key idea: A small request

If you only take one thing away from this course, please let it be this:
Probabilities are **numbers** and events are **sets**.
We can add up numbers (but not sets) and we can take unions and intersections of sets (but not numbers).

For the axioms to make sense, we can't just use any old event set $\mathcal{F}$. For one thing, we need $\Omega \in \mathcal{F}$; in fact all the events in **(A1-4)** need to be in $\mathcal{F}$. Our definition of a σ-algebra from the previous section gives us exactly the event set we need.

Key idea: Definition: probability space

If Ω is a set and $\mathcal{F}$ is a σ-algebra of subsets of Ω, and if $\mathbb{P}$ satisfies **(A1–4)** for events in $\mathcal{F}$, then the triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a *probability space*.

Try it out

Consider a finite sample space $\Omega = \{\omega_1, \dots, \omega_m\}$ of size $|\Omega| = m$. Then we can define a valid probability $\mathbb{P}$ by taking any numbers $p_1, \dots, p_m$ with $p_i \geq 0$ for all i and $\sum_{i=1}^{m} p_i = 1$ and declaring that for any event A,

$$\mathbb{P}(A) = \sum_{i:\omega_i \in A} p_i.$$

This satisfies the axioms **(A1–4)** (don't just believe me - check them for yourself).
By considering the event $A = \{\omega_i\}$, we see that $p_i = \mathbb{P}(\omega_i)$ is the probability of the elementary outcome ω_i.
In the simplest setting, we might assume that all the outcomes are *equally likely*, that is, $p_i = 1/m$ for all i. Note that in this case probability reduces to counting, since

$$\mathbb{P}(A) = \sum_{i:\omega_i \in A} \frac{1}{m} = \frac{|A|}{|\Omega|}.$$

As a concrete example, for tossing a *fair* die we would have $\Omega = \{1, 2, \dots, 6\}$, and $\mathbb{P}(A) = |A|/6$ so, for example,

$$\mathbb{P}(\text{score is odd}) = \mathbb{P}(\{1, 3, 5\}) = \frac{3}{6} = \frac{1}{2}.$$

We examine this setting in detail in Chapter 2.

Try it out

Consider a countably infinite sample space $\Omega = \{\omega_1, \omega_2, \dots\}$. Then we can define a valid probability $\mathbb{P}$ by taking any numbers $p_1, p_2, \dots$ with $p_i \geq 0$ for all i and $\sum_{i=1}^{\infty} p_i = 1$ and declaring that for any event A,

$$\mathbb{P}(A) = \sum_{i:\omega_i \in A} p_i.$$

This definition of a probability satisfies all of the axioms **(A1-A4)**.

For this course, we will usually assume that the probability distribution is given (and satisfies the axioms), without worrying too much about how the important practical task of finding the probabilities was carried out.

Textbook references

If you want more help with this section, check out:

- Section 1.6 in (Blitzstein and Hwang 2019);
- Section 1.1 in (Anderson, Seppäläinen, and Valkó 2018);
- or Section 1.3 in (Stirzaker 2003).

1.6 Consequences of the axioms

A host of useful results can be derived from A1–4.

Key idea: Consequences of the axioms

(C1) For any two events A and B,

$$\mathbb{P}(B\backslash A) = \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

(C2) For any event A, $\mathbb{P}(A^{\mathrm{c}}) = 1 - \mathbb{P}(A)$.
(C3) The probability of $\emptyset$ is $\mathbb{P}(\emptyset) = 0$.
(C4) For any event A, $\mathbb{P}(A) \le 1$.
(C5) If $A \subseteq B$ then $\mathbb{P}(A) \le \mathbb{P}(B)$ ("monotonicity").
(C6) For any two events A and B,

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

(C7) If $A_1, A_2, \dots, A_k$ are pairwise disjoint (so $A_i \cap A_j = \emptyset$ if $i \neq j$) then

$$\mathbb{P}\left(\bigcup_{i=1}^{k} A_i\right) = \sum_{i=1}^{k} \mathbb{P}(A_i).$$

(This property is called "finite additivity" in textbooks.)
(C8) For any events $A_1, A_2, \dots$, (these need not be pairwise disjoint),

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

(This one is sometimes referred to as "Boole's inequality.")
(C9) If $A_1 \subseteq A_2 \subseteq \cdots$ is an *increasing sequence* of events, then

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mathbb{P}(A_n).$$

If $A_1 \supseteq A_2 \supseteq \cdots$ is a *decreasing sequence* of events, then

$$\mathbb{P}\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mathbb{P}(A_n).$$

(This property is a bit more sophisticated than the previous ones. It establishes the "continuity of probability along monotone limits:" we can take limits, as long as the events in question form a monotone sequence. It will be really important in Probability II.)

Just one more consequence to go! Before we get there, we need the following simple but extremely useful idea: partitions.

Key idea: Definition: partition

We say that the events $E_1, E_2, \dots, E_k \in \mathcal{F}$ form a (finite) *partition* of the sample space Ω if:

i. they all have positive probability, i.e., $\mathbb{P}(E_i) > 0$ for all i;

ii. they are *pairwise disjoint*, i.e., $E_i \cap E_j = \emptyset$ whenever $i \neq j$; and
iii. their union is the whole sample space: $\cup_{i=1}^{k} E_i = \Omega$.

The definition extends to countably infinite partitions. We say that $E_1, E_2, \ldots \in \mathcal{F}$ form an infinite partition of Ω if:

i. $\mathbb{P}(E_i) > 0$ for all i;
ii. $E_i \cap E_j = \emptyset$ whenever $i \neq j$; and
iii. $\cup_{i=1}^{\infty} E_i = \Omega$.

For example, consider the sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$. Some partitions are:

$$\begin{aligned}
\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\} \\
\{1, 2\}, \{3, 4\}, \{5, 6\} \\
\{1, 2, 3\}, \{4, 5, 6\} \\
\{1\}, \{2, 3\}, \{4, 5, 6\} \\
\{1, 2, 3, 4, 5, 6\}
\end{aligned}$$

and so on.

Key idea

(C10) If E_1, E_2, ..., E_k form a partition then

$$\sum_{i=1}^{k} \mathbb{P}(E_i) = 1.$$

These consequences have an enormous effect on the way we work with probability. In particular, it turns out that we can solve most problems without ever having to explicitly write down the outcomes in our sample space, as in the next example. In fact, some people do probability without even defining a sample space.

Try it out

Jimmy's die has the numbers 2,2,2,2,5,5. Your die has numbers 1,1,4,4,4,4. You both throw and the highest number wins. Assuming all outcomes are equally likely, what is the probability that Jimmy wins?

Answer:

The event, J, that Jimmy wins happens if either Jimmy throws a 5 (call this event F), or if you throw a 1 (call this event A). Therefore $J = A \cup F$ and by C6,

$$\mathbb{P}(J) = \mathbb{P}(A) + \mathbb{P}(F) - \mathbb{P}(A \cap F).$$

As $\mathbb{P}(F) = 1/3$, $\mathbb{P}(A) = 1/3$ and $\mathbb{P}(A \cap F) = 4/36 = 1/9$ (by counting equally likely outcomes) we have

$$\mathbb{P}(J) = 1/3 + 1/3 - 1/9 = 5/9.$$

Finite sample spaces are a great way to build up our intuition for probability calculations. However, it is surprisingly easy to end up in a situation where things start to get complicated.

Try it out

What is the probability that, in an indefinitely long sequence of tosses of a fair coin, we will eventually see heads?
Answer:
The sample space Ω is infinite and consists of all sequences $\omega = (\omega_1, \omega_2, \ldots)$ with $\omega_i \in \{\mathrm{H}, \mathrm{T}\}$.
What is $\mathbb{P}$? Well, it certainly would be desirable that if we restrict to just a finite sequence of n tosses, then each of the 2^n possible outcomes (sequences) should be equally likely. It is a special case of a general theorem that such a $\mathbb{P}$ exists and is unique.
Now, let $A = \{\mathrm{H} \text{ occurs}\}$. Then the only way A can *not* occur is if there are *no* heads, i.e., $A^c = \{(\mathrm{TTT}\cdots)\}$. This is a single sequence, out of infinitely many, and it is intuitively clear that it should have probability 0. To prove this, it is enough to observe that $A^c \subseteq \{\text{first } n \text{ tosses T}\}$, so by monotonicity **(C5)**,

$$\mathbb{P}(A^c) \leq \mathbb{P}(\{\text{first } n \text{ tosses are T}\}) \leq 2^{-n}.$$

But this is true for any n, so we must have $\mathbb{P}(A^c) = 0$.
Another way to see this is as follows. Consider events defined for $n = 1, 2, \ldots$ by

$$A_n = \{\text{first H occurs on toss } n\} = \{\omega : \omega_k = \mathrm{T}, \text{ for all } k < n,\ \omega_n = \mathrm{H}\}.$$

This means that A_1 consists of sequences H⋯, A_2 consists of sequences TH⋯, and so on.
Now the event we are interested in is $A = \cup_{n=1}^{\infty} A_n$. So, by **(A4)**,

$$\mathbb{P}(A) = \sum_{n=1}^{\infty} \mathbb{P}(A_n) = \sum_{n=1}^{\infty} 2^{-n} = 1.$$

Note that a similar argument works if the coin is biased with probability $p \in (0, 1)$ of heads.

Advanced content

In fact, the sequence space Ω in the previous example is not even countable! To see this, a sequence $(d_1, d_2, \ldots)$ with each $d_i \in \{0, 1\}$ is called a *dyadic expansion* of $x \in [0, 1]$ if $x = \sum_{i=1}^{\infty} 2^{-i} d_i$. For example, $(1, 0, 0, \ldots)$ is a dyadic expansion of $1/2$, $(1, 1, 0, 0, \ldots)$ is $3/4$, and so on. The map between x and $(d_1, d_2, \ldots)$ is almost a bijection. It is not a bijection because of possible non-uniqueness of the dyadic expansion: e.g. $(0, 1, 1, 1, \ldots)$ is another expansion of $1/2$. It turns out that this problem only occurs for rational x, and can be circumvented. Thus we have (essentially) a bijection between $[0, 1]$ and the space of infinite sequences of 0s and 1s, which is another name for our coin tossing space Ω. This shows that Ω is uncountable.
It is remarkable that the probability $\mathbb{P}$ on infinite sequences of coin tosses turns out to correspond (under the bijection by dyadic expansion) to nothing other than the *uniform* distribution on $[0, 1]$, that is the measure defined by lengths of intervals. This is the famous *Lebesgue* measure.

Textbook references

If you want more help with this section, check out:

- Section 1.6 in (Blitzstein and Hwang 2019);
- Section 1.4 in (Anderson, Seppäläinen, and Valkó 2018);

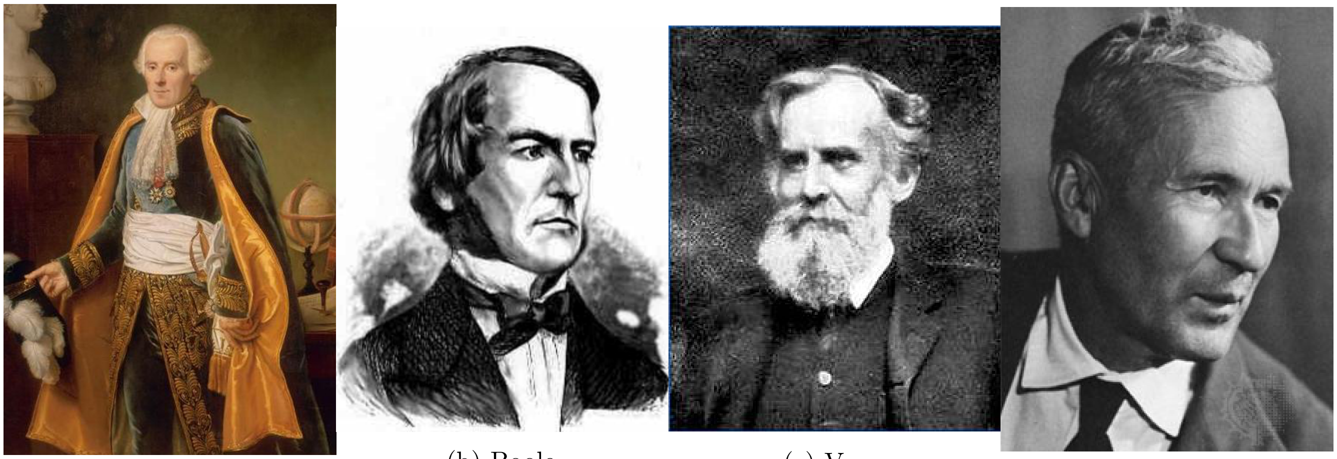
- or Sections 1.4 and 1.5 in (Stirzaker 2003).

1.7 Historical context

Sets are important not only for probability theory, but for all of mathematics. In fact, all of standard mathematics can be formulated in terms of set theory, under the assumption that sets satisfy the ZFC axioms; see for instance this Wikipedia page.

The foundations of probability have a long and interesting history (Hacking 2006; Todhunter 2014). The classical theory owes much to Pierre-Simon Laplace (1749–1827): see (Laplace 1825). However, a rigorous mathematical foundation for the theory was lacking, and was posed as part of one of David Hilbert's (1862–1943) famous list of problems in 1900 (the 5th problem). After important work by Henri Lebesgue (1875–1941) and 'Emile Borel (1871–1956), it was Andrey Kolmogorov who succeeded in 1933 in providing the axioms that we use today (see the 1950 edition of his book (Kolmogorov 1950)). This approach declares that probabilities are *measures*.

A measure μ can be defined on any set Ω with a σ-algebra of subsets $\mathscr{F}$, and the defining axioms are versions of A1 and A4. The special property of a probability measure is just that $\mu(\Omega) = 1$. Measure theory is the theory that gives mathematical foundation to the concepts of length, area, and volume. For example, on $\mathbb{R}$ the unique measure that has $\mu(a, b) = b - a$ for intervals (a, b) is the *Lebesgue measure*.



(a) Laplace

(b) Boole

(c) Venn

(d) Kolmogorov

Figure 1.2: Laplace, Boole, Venn, and Kolmogorov

George Boole (1815–1864) and John Venn (1834–1923) both wrote books concerned with probability theory (Boole 1854), (Venn 1888); both were working before the formulation of Kolmogorov's axioms.

As mentioned in Section 1.4, it is necessary in the general theory of probability to restricting events to some σ-algebra. The reason for this is that in standard ZFC set theory, when Ω is uncountable (such as $\Omega = [0, 1]$ the unit interval), it follows from an argument by Vitali (1905) that many natural probability assessments, such as the continuous uniform distribution, cannot be modelled by a probability defined on *all* subsets of Ω satisfying A1–4: see for instance Chapter 1 of (Rosenthal 2007). In the case where Ω is countable, one can always define $\mathbb{P}$ on the whole of 2^{Ω}. In the case where Ω is uncountable, we usually do not explicitly mention Ω at all (when we work with continuous random variables, for example).

The formulation of the infinite coin-tossing experiment in Section 1.6 leads to the connection between coin tossing and the Lebesgue measure, as first described by Hugo Steinhaus in a 1923 paper.

An alternative approach to probability theory is to do away with axiom A4, in which case some of these technical issues can be avoided, at the expense of certain pathologies; however, in the standard approach to modern probability, based on *measure theory*, A4 is a central part of the theory.

2 Equally likely outcomes and counting principles

Goals

1. Understand the equally likely outcomes model of classical probability.
2. Know counting principles, and when and how to apply them on specific problems.

In Chapter 1 we have seen the abstract formulation of probability theory; next we turn to the question of how the probabilities themselves may be assigned.

The most basic scenario occurs when our experiment has a finite number of possible outcomes which we deem to be *equally likely*.

Such situations rarely—not to say never—occur in practice, but serve as good models in extremely controlled environments such as in gambling or games of chance. However, this situation (which will essentially come down to counting) gives us a good initial setting in which to obtain some very useful insights into the nature and calculus of probability.

2.1 Classical probability

Suppose that we have a finite sample space Ω. Since Ω is finite, we can list it as a collection of $m = |\Omega|$ possible outcomes:

$$\Omega = \{\omega_1, \dots, \omega_m\}.$$

In the equally likely outcomes model (also sometimes known as classical probability) we suppose that each outcome has the same probability:

$$\mathbb{P}(\omega) = \frac{1}{|\Omega|} \text{ for each } \omega \in \Omega,$$

or, in the notation above, $\mathbb{P}(\omega_i) = 1/m$ for each i.

The axioms of probability then allow us to determine the probability of any event $A \subseteq \Omega$: by **C7**,

$$\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\omega) = \frac{|A|}{|\Omega|} \text{ for any event } A \subseteq \Omega.$$

This is a particular case of the discrete sample space discussed in Chapter 1.

Definition: Equally likely outcomes

Consider a scenario with m equally likely outcomes enumerated as $\Omega = \{\omega_1, \dots, \omega_m\}$. In the equally likely outcomes model, the probability of an event $A \subseteq \Omega$ is declared to be

$$\mathbb{P}(A) := \frac{|A|}{|\Omega|}.$$

Using this definition, we meet all of the axioms **(A1–A4)** (checking each of them comes down to what we know about counting). Remember that in the case of a finite state space, we always have the option to take $\mathcal{F} = 2^\Omega$ as our σ-algebra.

Examples

1. Draw a card at random from a well-shuffled pack, so that each of the 52 cards is equally likely to be chosen. Typical events are that the card is a spade (a set of 13 outcomes), the card is a queen (a set of 4 outcomes), the card is the queen of spades (a set of a single outcome). In the equally likely outcomes model, the probability of drawing the queen of spades (or any other specified card) is $1/52$, the probability of drawing a spade is $13/52$, and the probability of drawing a queen is $4/52$.

2. Flip a coin and see whether it falls heads or tails, each assumed equally likely; then 'heads' or 'tails' each has probability $1/2$.

3. Roll a fair cubic die to get a number from 1 to 6. Here the word 'fair' is used to mean each outcome is equally likely. Then $\Omega = \{1, \dots, 6\}$ and $\mathbb{P}(A) = |A|/6$. For example, if $A_1 = \{2\}$ (the score is 2) we get $\mathbb{P}(A_1) = 1/6$, while if $A_2 = \{1, 3, 5\}$ (the score is odd) we get $\mathbb{P}(A_2) = 3/6 = 1/2$.

4. If we roll a pair of fair dice then outcomes are pairs (i, j) so there are 36 possible outcomes. If we assume that the outcomes are equally likely, then the probability of getting a pair of 6's is $1/36$, for example.

The classical interpretation of probability is the most straightforward approach we can take, just as counting can be seen as "basic" mathematics. It is a good place to start and there are many important situations where intuitively it seems reasonable to say that each outcome of a particular collection is equally likely.

To extend the theory or apply it in practice we have to address situations where there are no candidates for equally likely outcomes or where there are infinitely many possible outcomes and work out how to find probabilities to put into calculations that give useful predictions. We will come back to some of these issues later; but bear in mind that however we come up with our probability model, the same system of axioms that we saw in Chapter 1 applies.

Textbook references

If you want more help with this section, check out:

- Section 1.3 in (Blitzstein and Hwang 2019);
- or Section 1.2 in (Anderson, Seppäläinen, and Valkó 2018).

2.2 Counting principles

Given a finite sample space and assuming that outcomes are equally likely, to determine probabilities of certain events comes down to counting.

For example, in drawing a poker hand of five cards from a well-shuffled deck of 52 cards, the probability of having a 'full house' (meaning two cards of one denomination and three of another, e.g., two Kings and three 7s) is given by the number of hands that are full houses divided by the total number of hands (each hand being equally likely).

These counting problems need careful thought, and we will describe some counting principles for some of the most common situations. There is some common ground with the *Discrete Maths* course; here we have a slightly different emphasis.

2.2.1 The multiplication principle

Counting principle: Multiplication

Suppose that we must make k choices in succession where there are:

- m_1 possibilities for the first choice,
- m_2 possibilities for the second choice,
- $\vdots$
- m_k possibilities for the kth choice,

and the number of possibilities at each stage does not depend on the outcomes of any previous choices. The total number of distinct possible selections is

$$m_1 \times m_2 \times m_3 \times \cdots \times m_k = \prod_{i=1}^{k} m_i.$$

For instance, in a standard deck of playing cards, each card has a *denomination* and a *suit*. There are 13 possible denominations: A(ce), 2, 3, ..., 10, J(ack), Q(ueen), K(ing). There are 4 possible suits: ♡ (heart), ♢ (diamond), ♣ (club), ♠ (spade). Because all combinations of denomination and suit are allowed, the multiplication principle applies: there are $13 \times 4 = 52$ cards in a standard deck.

We will see many applications of counting to dealing cards from a well-shuffled deck. Counting the possibilities is affected by (i) whether the *order* of dealing is important, and (ii) how we *distinguish* the cards: e.g. we may only be interested in their colour (so all red cards are the same) or their suit or their denomination.

Examples

1. A hotel serves 3 choices of breakfast, 4 choices of lunch and 5 choices of dinner so a guest selects from $3 \times 4 \times 5$ different combinations of the three meals (assuming we opt to have all three).
2. A coffee bar has 5 different papers to choose from, 19 types of coffee and 7 different snacks. This means there are $6 \times 20 \times 8 = 960$ distinct selections of coffee, snack and paper. Of these 5 involve no coffee or snack (which the staff may object to) plus one has no coffee, snack or paper!
3. PINs are made up of 4 digits (0–9) with the exceptions that (i) they cannot be four repetitions of a single digit; (ii) they cannot form increasing or decreasing consecutive sequences, e.g. 3456 and 8765 are excluded. How many possible four-digit PINs are there?

   Ignoring restrictions there are $10^4 = 10,000$ distinct PINs. There are 10 PINs with the same digit repeated, namely 0000, 1111, ..., 9999. Increasing sequences start with $0, 1, 2, \ldots, 6$ and decreasing sequences start with $9, 8, \ldots, 3$, so there are seven options for each. This leaves $10,000 - 24 = 9,976$ permitted PINs.

All of the following counting principles are effectively consequences of the multiplication principle.

2.2.2 Order matters; objects are distinct

First, we look at ordered choices of distinct objects. In this case, we distinguish between *selection with replacement*, where the same object can be selected multiple times, and *selection without replacement*, where each object can only be selected at most once.

Counting principle: Selection with replacement for ordered choices

Suppose that we have a collection of m distinct objects and we select r of them with replacement. The number of different ordered lists (ordered r-tuples) is

$$\underbrace{m \times \cdots \times m}_{r \text{ times}} = m^r.$$

Counting principle: Selection without replacement for ordered choices

Suppose that we have a collection of m distinct objects and we select $r \leq m$ of them without replacement. The number of different ordered lists (ordered r-tuples) is

$$(m)_r := \underbrace{m \times (m-1) \times (m-2) \times \cdots \times (m-r+1)}_{r \text{ terms}} = \frac{m!}{(m-r)!}.$$

The *falling factorial* notation $(m)_r$ (sometimes also denoted $m^{\underline{r}}$) is simply a convenient way to write $\frac{m!}{(m-r)!}$. In the special case where $r = m$ we set $0! = 1$ and then $(m)_m = m!$ is the number of permutations of the m objects. If m is large, and r is much smaller than m, then $(m)_r \approx m^r$.

Example

The number of ways we can deal out four cards in order from a pack of cards is $(52)_4$ and the number of ways we can arrange the four aces in order is 4! so the probability of finding the four aces on top of a well-shuffled deck is

$$\frac{4!}{(52)_4} = \frac{4 \times 3 \times 2 \times 1}{52 \times 51 \times 50 \times 49}.$$

This probability is approximately 3.7×10^{-6} or about 1 in 270,000.

Try it out

There are $n < 365$ people in a room. Let B be the event that (at least) two of them have the same birthday. (We ignore leap years.)
What is $\mathbb{P}(B)$? How big must n be so that $\mathbb{P}(B) > 1/2$?
Answer:

Here the equally likely outcomes are the ordered length-n lists of possible birthdays:

$$(\text{person 1's birthday}, \text{person 2's birthday}, \dots, \text{person } n\text{'s birthday}).$$

The number of possible outcomes is

$$365 \times 365 \times \cdots \times 365 = 365^n.$$

This is the denominator in our probability.
For the numerator, we must work out how many outcomes are in B. In fact, it is easier to count outcomes in B^{c}, where everyone has a different birthday. There are

$$365 \times 364 \times \cdots \times (365 - n + 1) = (365)_n$$

of these. So

$$\mathbb{P}(B) = 1 - \mathbb{P}(B^{\mathrm{c}}) = 1 - \frac{(365)_n}{365^n}.$$

It turns out that $\mathbb{P}(B) \approx 1/2$ for $n = 23$.

Advanced content

Here's one way to get this. Note that

$$\frac{(365)_n}{365^n} = 1 \times \left(1 - \frac{1}{365}\right) \times \left(1 - \frac{2}{365}\right) \times \cdots \times \left(1 - \frac{n-1}{365}\right).$$

Now $1 - x \leq e^{-x}$, and in fact the inequality is very close to equality for $x = 1/365$, being close to zero. In any case,

$$\begin{aligned} \frac{(365)_n}{365^n} &\leq e^{-x}\, e^{-2x}\, e^{-3x} \cdots e^{-(n-1)x} \\ &= \exp\{-(1 + 2 + \cdots + n - 1)x\} \\ &= \exp\left\{-\frac{(n-1)n}{2 \times 365}\right\}. \end{aligned}$$

2.2.3 Order doesn't matter; objects are distinct

In this section, we move on to think about the scenario where the order in which objects are selected *doesn't* matter. This can arise in situations such as dealing a hand of cards, or separating a class into two teams.

Counting principle: Selection without replacement for unordered choices

Suppose that we have a collection of m distinct objects and we select a subset of $r \leq m$ of them without replacement. The number of distinct subsets of size r is

$$\binom{m}{r} := \frac{(m)_r}{r!} = \frac{m!}{r!\,(m-r)!}.$$

To see this, first count the number of distinct ordered lists of r objects—this is $(m)_r$. Each unordered subset has been counted $(r)_r = r!$ times as this is the number of distinct ways of arranging r different objects. Therefore the $(m)_r$ ordered selections can be grouped into collections of size $r!$, each representing a particular subset, and the result follows by dividing.

The expression $\binom{m}{r}$ is the *binomial coefficient* for choosing r objects from m and is often called 'm-choose-r'. Note that

$$\binom{m}{r} = \binom{m}{m-r}$$

as we can choose to take r objects from m in exactly the same number of ways that we can choose to leave behind r objects i.e., take $m - r$ objects.

Try it out

What is the probability of finding no aces in a four-card hand dealt from a well-shuffled deck?
Answer:
Let's answer this by treating hands as unordered selections. Then there are

$$\binom{52}{4} = \frac{52 \times 51 \times 50 \times 49}{4 \times 3 \times 2 \times 1} = 270,725$$

distinct unordered hands of four cards. The number of these with no aces is

$$\binom{48}{4} = \frac{48 \times 47 \times 46 \times 45}{4 \times 3 \times 2 \times 1} = 194,580,$$

and so the probability of finding no aces in a four card hand is

$$\binom{48}{4} \Big/ \binom{52}{4} = \frac{48 \times 47 \times 46 \times 45}{52 \times 51 \times 50 \times 49} \approx 0.7187.$$

Alternatively, we could answer this by treating the hands as *ordered* selections (the order corresponding to the order of the deal, say). Of course, this will give different numerator and denominator in our calculation, but the final answer must be the same! As ordered selections, there are

$$(52)_4 = 52 \times 51 \times 50 \times 49$$

distinct hands. The number of these with no ace is

$$(48)_4 = 48 \times 47 \times 46 \times 45.$$

Our probability is then $(48)_4/(52)_4$ which is the same as before.

In this simple example, either method is relatively straightforward, but in many examples, it is much more natural to treat the selections as ordered/undordered. For hands of cards, treating them as unordered selections usually works best. For something like rolling dice, it usually makes sense to treat them as ordered selections.

Try it out

You are dealt five cards from a well-shuffled deck. Let A be the event that exactly four cards are of the same suit. What is $\mathbb{P}(A)$?

Answer:

There are $\binom{52}{5}$ different unordered selections for the hand, and all are equally likely. How many of these unordered selections are in A? We need to describe a subset of 5 elements such that exactly 4 have the same suit. We build this up sequentially:

- We first choose the suit that we are going to use for the four cards: 4 possibilities.
- Then we choose the four denominations (unordered) for those cards: $\binom{13}{4}$ possibilities.
- All that remains is to choose the last card, which must be of a different suit than the four already chosen: $3 \times 13 = 39$ possibilities.

So the answer is

$$\mathbb{P}(A) = \frac{4 \times \binom{13}{4} \times 39}{\binom{52}{5}} \approx 0.0429.$$

Try it out

In 'Lotto Extra' you have to select 6 numbers from 1 to 49. You win the big prize if 6 randomly drawn numbers match your selection. Let W be the event that you win. Let M_4 be the event that you match exactly 4 out of 6 numbers. Find the probabilities of W and M_4.

Answer:

We model the outcomes of the Lotto draw as unordered selections, so there are $\binom{49}{6} = 13,983,816$ outcomes in total. The event W contains only one of them (your entry)! So $\mathbb{P}(W) = 1/13,983,816$. Now M_4 uses any 4 of your numbers plus any 2 of the remaining $49 - 6 = 43$ numbers. So the number of outcomes in M_4 is

$$\binom{6}{4} \times \binom{43}{2} = \frac{6 \times 5}{2 \times 1} \times \frac{43 \times 42}{2 \times 1} = 15 \times 43 \times 21 = 13,545.$$

Then $\mathbb{P}(M_4) = 13,545/13,983,816 \approx 0.001$.

Advanced content

The same counting arguments can be used when we need to divide m objects into $k > 2$ groups: arranging m distinguishable objects into k groups with sizes $r_1, \dots, r_k$ where $r_1 + \cdots + r_k = m$ can be done in

$$\binom{m}{r_1, \dots, r_k} := \frac{m!}{r_1! \cdots r_k!}$$

ways. The expression $\binom{m}{r_1, \dots, r_k}$ is called the *multinomial coefficient* (Anderson, Seppäläinen, and Valkó 2018 Example 6.7).

2.2.4 Separating objects into groups

In the final section of this chapter, we look into how we can *group* objects: either by combining different types of object into one big group, or by separating a big group into smaller ones.

Counting principle: Two types of object

Suppose that we have m objects, r of type 1 and $m-r$ of type 2, where objects are indistinguishable from others of their type. The number of distinct, ordered choices of the m objects is

$$\binom{m}{r}.$$

For example, suppose we have four red tokens, and three black ones. Then there are $7!/(4!\,3!) = 35$ different ways to lay them out in a row. The probability that they will be alternately red and black is $1/35$ as there is only one such ordering.

To see why, note that each distinct order for laying out all of the token in a row is precisely the same as choosing 4 of the 7 positions for red ones. In other words, it is an unordered choice of 4 positions from the 7 distinct positions.

Try it out

A coin is tossed 7 times. Let E be the event that a total of 3 heads is obtained. What is $\mathbb{P}(E)$?
Answer:
Consider ordered sequences of H and T: then there are $2^7 = 128$ possible sequences, e.g. HTHTHTT. How many of them are in E? We choose the 3 places where H occurs: $\binom{7}{3} = 35$ ways to do this. The other places are taken by Ts. So the answer is

$$\mathbb{P}(E) = \frac{35}{128}.$$

Advanced content

More generally, using the positions argument again, the multinomial coefficient is the number of ordered choices of objects with k types, r_i of type i, which are indistinguishable within each type.

Counting principle: Separating into groups

The number of ways to divide m indistinguishable objects into k distinct groups is

$$\binom{m+k-1}{m} = \binom{m+k-1}{k-1}.$$

This counting principle lets us work out how many different ways there are to divide one group into smaller groups. My favourite example is a packet of Skittles: if there are 16 or 17 of them in a bag, how many different combinations of the five different flavours could we have?

We can count the number of choices with the ‘sheep-and-fences’ method. Placing all the objects in a line, separated into their groups, there are $k-1$ “fenceposts” between the k groups of sheep (or Skittles).

For example, with 6 objects in 4 groups, we could represent "three in group A, one in group B, none in group C and two in group D" with the drawing $* * * \mid * \mid\mid **$.

We draw $m+k-1$ 'things' in total (stars and fences). This means that the number of groupings of the objects is the same as the number of choices for the locations of the $k-1$ fences among the $m+k-1$ 'things', or $\binom{m+k-1}{k-1} = \binom{m+k-1}{m}$.

Textbook references

If you want more help with this section, check out:

- Section 1.4 in (Blitzstein and Hwang 2019);
- Appendix C in (Anderson, Seppäläinen, and Valkó 2018);
- or Chapter 3 in (Stirzaker 2003).

2.3 Historical context

Classical probability theory originated in calculation of odds for games of chance; as well as contributions by Pierre de Fermat (1601–1665) and Blaise Pascal (1623–1662), comprehensive approaches were given by Abraham de Moivre (1667–1754) (Moivre 1756), Laplace (1749–1827) (Laplace 1825), and Sim'eon Poisson (1781–1840). A collection of these classical methods made just before the advent of the modern axiomatic theory can be found in (Whitworth 1901).

3 Conditional probability and independence

Goals

1. Know the definition of conditional probability and its properties
2. Have a solid knowledge of the partition theorem and Bayes' theorem, recognizing situations where one can apply them.
3. Understand the concept of independence.

3.1 Conditional probability

Definition: conditional probability

For events $A, B \subseteq \Omega$, the *conditional probability* of A *given* B is

$$\mathbb{P}(A \mid B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \qquad \text{whenever} \;\; \mathbb{P}(B) > 0.$$

In this course, when $\mathbb{P}(B) = 0$, $\mathbb{P}(A \mid B)$ is undefined. The usual interpretation is that $\mathbb{P}(A \mid B)$ represents our probability for A after we have observed B. Conditional probability is therefore very important for statistical reasoning, for example:

- In legal trials. How can we use DNA (or other) evidence to determine the chance that an accused person is guilty?
- Medical screening. How can we make best use of the information from large scale cancer screening programs?

Unfortunately, conditional probability is not always well understood. There are several well-known legal cases that have involved a serious error in probabilistic reasoning: see e.g. Example 2.4.5 of (Anderson, Seppäläinen, and Valkó 2018).

For example, if we roll a fair six-sided die, the conditional probability that the score is odd, given that the score is at most 3, is

$$\mathbb{P}(\text{odd} \mid \text{ at most } 3) = \frac{\mathbb{P}(\{1,3\})}{\mathbb{P}(\{1,2,3\})} = \frac{2/6}{3/6} = \frac{2}{3}.$$

Try it out

Throw three fair coins. What is the conditional probability of at least one head (event A) given at least one tail (event B)?
Answer:
Let H be the event 'all heads', T the event 'all tails'. Then $\mathbb{P}(B) = 1 - \mathbb{P}(H) = 7/8$ and $\mathbb{P}(A \cap B) =$

$1 - \mathbb{P}(H) - \mathbb{P}(T) = 6/8$ so that

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{6/8}{7/8} = \frac{6}{7}.$$

Try it out

Consider a family with two children, whose sex we do not know. The possible sexes are listed by the sample space $\Omega = \{\mathrm{BB}, \mathrm{BG}, \mathrm{GB}, \mathrm{GG}\}$, with the eldest first. Assume that all outcomes are equally likely. Consider the events

$$\begin{aligned} A_1 &= \{\mathrm{GG}\} = \{\text{both girls}\}, \\ A_2 &= \{\mathrm{GB}, \mathrm{BG}, \mathrm{GG}\} = \{\text{at least one girl}\}, \\ A_3 &= \{\mathrm{GB}, \mathrm{GG}\} = \{\text{first child is a girl}\}. \end{aligned}$$

Find $\mathbb{P}(A_1 \mid A_2)$, $\mathbb{P}(A_2 \mid A_1)$, and $\mathbb{P}(A_1 \mid A_3)$.
Answer:
We compute

$$\begin{aligned} \mathbb{P}(A_1 \mid A_2) &= \frac{\mathbb{P}(A_1 \cap A_2)}{\mathbb{P}(A_2)} = \frac{\mathbb{P}(\{\mathrm{GG}\})}{\mathbb{P}(\{\mathrm{GB}, \mathrm{BG}, \mathrm{GG}\})} \\ &= \frac{1/4}{3/4} = \frac{1}{3}. \end{aligned}$$

Similarly,

$$\mathbb{P}(A_2 \mid A_1) = \frac{\mathbb{P}(A_1 \cap A_2)}{\mathbb{P}(A_1)} = \frac{\mathbb{P}(\{\mathrm{GG}\})}{\mathbb{P}(\{\mathrm{GG}\})} = 1,$$

and

$$\mathbb{P}(A_1 \mid A_3) = \frac{\mathbb{P}(A_1 \cap A_3)}{\mathbb{P}(A_3)} = \frac{\mathbb{P}(\{\mathrm{GG}\})}{\mathbb{P}(\{\mathrm{GB}, \mathrm{GG}\})} = \frac{1/4}{2/4} = \frac{1}{2}.$$

Try it out

Consider throwing two standard dice. Consider the events $F =$ first die shows 6, and $T =$ total is 10. Calculate $\mathbb{P}(F)$ and $\mathbb{P}(F \mid T)$. Before doing any calculation, do you expect $\mathbb{P}(F \mid T)$ to be higher or lower than $\mathbb{P}(F)$? (Hint: 10 is a high total. We'll see later that the 'average' total score on two dice is 7.)
Answer:
The possible outcomes are ordered pairs of the numbers 1 to 6, so $|\Omega| = 6^2 = 36$. In F are all outcomes of the form $(6, ?)$. There are 6 of those, so $\mathbb{P}(F) = 6/36 = 1/6$.
Now $T = \{(6,4), (5,5), (4,6)\}$ so $F \cap T = \{(6,4)\}$, and $\mathbb{P}(F \mid T) = (1/36)/(3/36) = 1/3 > 1/6$.
Similarly, if the total had been 5 we would know that F was impossible!

Textbook references

If you want more help with this section, check out:

- Section 2.2 in (Blitzstein and Hwang 2019);
- Section 2.1 in (Anderson, Seppäläinen, and Valkó 2018);

- or Section 2.1 in (Stirzaker 2003).

3.2 Properties of conditional probability

In this section, we'll meet five key properties of conditional probability.

Key idea: properties of conditional probability

(P1) For any event $B \subseteq \Omega$ for which $\mathbb{P}(B) > 0$, $\mathbb{P}(\cdot \mid B)$ satisfies axioms **A1–A4** (i.e., is a probability on Ω) and therefore also satisfies **C1–C10**.

For example, **C6** for conditional probabilities says that, if $\mathbb{P}(C) > 0$,

$$\mathbb{P}(A \cup B \mid C) = \mathbb{P}(A \mid C) + \mathbb{P}(B \mid C) - \mathbb{P}(A \cap B \mid C).$$

Key idea: properties of conditional probability: multiplication

(P2) For any events A and B with $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$,

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

More generally, for any A, B, and C,

$$\mathbb{P}(A \cap B \mid C) = \mathbb{P}(B \mid C)\,\mathbb{P}(A \mid B \cap C), \qquad \text{if } \mathbb{P}(B \cap C) > 0. \tag{3.1}$$

Some people refer to **P2** as the multiplication rule for probabilities.

Both **P1** and **P2** can be deduced from the definition of probability. For example, Equation 3.1 follows from the fact that

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

Try it out

Derek is playing Dungarees & Dragons. He rolls an octahedral die to generate the occupant of the room he has just entered. He knows that with probability 3/8 it will be a Goblin, otherwise it will be a Hobbit. A Goblin has a 1 in 4 chance of being equipped with a spiky club. What is the chance that he encounters a Goblin with a spiky club?

Answer:

Let G be the event that the occupant is a Goblin, and let C be the event that the occupant has a spiky club. We are told that $\mathbb{P}(G) = 3/8$ and $\mathbb{P}(C \mid G) = 1/4$, so $\mathbb{P}(G \cap C) = \mathbb{P}(G)\mathbb{P}(C \mid G) = (3/8) \times (1/4) = 3/32$.

Our next property is a more general version of the multiplication rule.

Key idea: properties of conditional probability: multiplication (again)

(P3): For any events $A_0, A_1, \ldots, A_k$ with $\mathbb{P}\left(\cap_{i=0}^{k-1} A_i\right) > 0$,

$$\mathbb{P}\left(\bigcap_{i=1}^{k} A_i \mid A_0\right) = \mathbb{P}(A_1 \mid A_0) \times \mathbb{P}(A_2 \mid A_1 \cap A_0) \times \cdots \times \mathbb{P}\left(A_{k-1} \mid \bigcap_{i=0}^{k-2} A_i\right) \times \mathbb{P}\left(A_k \mid \bigcap_{i=0}^{k-1} A_i\right).$$

When $k = 2$, we get **P2**; for $k = 3$, this becomes

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

We can prove this by repeatedly applying Equation 3.1 (in this case, we use it twice).

Try it out

If Derek encounters a Goblin armed with a spiky club, the Goblin will attack, causing a wound with probability 1/2. A Goblin without a spiky club will flee. If Derek encounters a Hobbit, the Hobbit will offer him a cup of tea. What is the probability that Derek is wounded by this encounter?
Answer: Let W be the event that Derek is wounded. Then

$$\mathbb{P}W = \mathbb{P}(G \cap C \cap W) = \mathbb{P}(G)\mathbb{P}(C \mid G)\mathbb{P}(W \mid C \cap G) = \frac{3}{8} \cdot \frac{1}{4} \cdot \frac{1}{2} = \frac{3}{64}.$$

Key idea: properties of conditional probability: partitions

(P4) If $E_1, E_2, \ldots, E_k$ form a partition then, for any event A, we have

$$\mathbb{P}(A) = \sum_{i=1}^{k} P(E_i)\,P(A \mid E_i)\,. \tag{3.2}$$

More generally, if $\mathbb{P}(B) > 0$,

$$\mathbb{P}(A \mid B) = \sum_{i=1}^{k} P(E_i \mid B)\,P(A \mid E_i \cap B)\,.$$

This result is often called the *partition theorem*, or the *law of total probability*. (If you've forgotten what a partition is, head back to Section 1.6.)

To prove **P4** is true, we first use **P2** on the right-hand side of Equation 3.2 to get

$$\sum_{i=1}^{k} P(E_i)\,P(A \mid E_i) = \sum_{i=1}^{k} P(A \cap E_i)\,.$$

But since the E_i form a partition, they are pairwise disjoint, and hence so are the $A \cap E_i$, so by **C7**

$$\sum_{i=1}^{k} P(A \cap E_i) = P(\cup_{i=1}^{k}(A \cap E_i)) = P(A \cap (\cup_{i=1}^{k} E_i))\,,$$

but since the E_i form a partition, $\cup_{i=1}^{k} E_i = \Omega$, giving the result. You should check that **P4** remains true (with $k = \infty$) for infinite partitions.

Try it out

Back in the land of Dungarees and Dragons, suppose that the Goblin player has a special token that he will play so that even unarmed Goblins will attack, rather than flee. An unarmed Goblin causes a wound with probability 1/6. Now what is the chance that Derek is wounded?

Answer:

The partition we use is G^{c}, $G{\cap}C$, and $G{\cap}C^{c}$. We know from our previous examples that $P(G^{c}) = 5/8$, $\mathbb{P}(G\cap C) = 3/32$, and $P(G\cap C^{c}) = \mathbb{P}(G)P(C^{c} \mid G) = 9/32$.

We also know that $P(W \mid G^{c}) = 0$, $P(W \mid G\cap C) = 1/2$, and, now, $P(W \mid G\cap C^{c}) = 1/6$. So

$$\begin{aligned}\mathbb{P}W &= P(G^{c})\,P(W \mid G^{c}) + \mathbb{P}(G\cap C)P(W \mid G\cap C) + P(G\cap C^{c})\,P(W \mid G\cap C^{c})\\ &= 0 + \frac{3}{32}\cdot\frac{1}{2} + \frac{9}{32}\cdot\frac{1}{6} = \frac{3}{32}.\end{aligned}$$

Try it out

Three machines, A, B and C, produce components. 10% of components from A are faulty, 20% of components from B are faulty and 30% of components from C are faulty. Equal numbers from each machine are collected in a packet. One component is selected at random from the packet. What is the probability that it is faulty?

Answer:

Let F be the event that the component is faulty. Let M_A, M_B, M_C be the events that the component is from machines A, B, C respectively. Then M_A, M_B, M_C form a partition so

$$\begin{aligned}P(F) &= P(M_A)\,P(F \mid M_A) + P(M_B)\,P(F \mid M_B) + P(M_C)\,P(F \mid M_C)\\ &= 0.1 \times \frac{1}{3} + 0.2 \times \frac{1}{3} + 0.3 \times \frac{1}{3} = 0.2.\end{aligned}$$

The most important result in conditional probability is *Bayes' theorem.* It allows us to express the conditional probability of an event A given B in terms of the “inverse” conditional probability of B given A.

Key idea: properties of conditional probability: Bayes theorem

(P5) For any events A and B with $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$,

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

More generally, if $\mathbb{P}(A \mid C) > 0$ and $\mathbb{P}(B \mid C) > 0$, then

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

Try it out

Suppose that in the previous example, the component was indeed faulty. What is the probability that it came from machine A?

Answer:

We have, by Bayes' theorem (**P5**),

$$P(M_A \mid F) = \frac{P(F \mid M_A)\, P(M_A)}{P(F)} = \frac{(1/10)(1/3)}{1/5} = \frac{1}{6}.$$

Try it out

Pat ends up in the pub of an evening with probability 3/10. If she goes to the pub, she will get drunk with probability 1/2. If she stays in, she will get drunk with probability 1/5. What is the probability that she gets drunk? Given that she does get drunk, what is the probability that she went to the pub?

Answer:

Let P be the event that she goes to the pub, and D be the event that she gets drunk. Then we are told that $P(P) = 3/10$, $P(D \mid P) = 1/2$ and $P(D \mid P^c) = 1/5$. Using the partition P, P^c we get

$$\mathbb{P}(D) = P(P)\, P(D \mid P) + P(P^c)\, P(D \mid P^c) = \frac{3}{10} \cdot \frac{1}{2} + \frac{7}{10} \cdot \frac{1}{5} = \frac{29}{100}.$$

Then, by Bayes' theorem (P5),

$$P(P \mid D) = \frac{P(D \mid P)\, P(P)}{\mathbb{P}(D)} = \frac{\frac{3}{10} \cdot \frac{1}{2}}{\frac{29}{100}} = \frac{15}{29}.$$

Try it out

There are three regions (A,B,C) in a country with populations in relative proportions $5:3:2$. In region A, 5% of people own a rabbit. In region B, it is 10%, and in region C, it is 15%.
i.What proportion of people nationally own rabbits? ii. What proportion of rabbit-owners come from region A?

Answer:

Let A, B, C be the events that a randomly-chosen individual comes from regions A, B, C respectively. Let R be the event that the individual is a rabbit owner. Then

$$P(R) = \mathbb{P}(A)P(R \mid A) + \mathbb{P}(B)P(R \mid B) + \mathbb{P}(C)P(R \mid C) = \frac{5}{10} \cdot \frac{1}{20} + \frac{3}{10} \cdot \frac{1}{10} + \frac{2}{10} \cdot \frac{3}{20} = \frac{17}{200}.$$

And, by Bayes' theorem,

$$P(A \mid R) = \frac{P(R \mid A)\, \mathbb{P}(A)}{P(R)} = \frac{\frac{5}{10} \cdot \frac{1}{20}}{\frac{17}{200}} = \frac{5}{17}.$$

Try it out

One of a set of n people committed a crime. A suspect has been arrested, and DNA evidence is a match. Consider the events $G =$ suspect is guilty, and $E =$ DNA evidence is a match. Suppose that we initially believe that $\mathbb{P}(G) = \alpha/n$. The probability of a 'false positive' DNA match is $P(E \mid G^c) = p$.
What is our new probability that the suspect is guilty, given the DNA evidence?

Answer:

We use the partition G, G^c. Then, by P4,

$$\begin{aligned} P(E) &= P(E \mid G)\,\mathbb{P}(G) + P(E \mid G^c)\,P(G^c) \\ &= 1 \times \frac{\alpha}{n} + p \times \left(1 - \frac{\alpha}{n}\right) \\ &= \frac{\alpha + (n-\alpha)p}{n}. \end{aligned}$$

Then by Bayes' theorem (P5),

$$P(G \mid E) = \frac{P(E \mid G)\,\mathbb{P}(G)}{P(E)} = \frac{\alpha/n}{(\alpha + (n-\alpha)p)/n} = \frac{\alpha}{\alpha + (n-\alpha)p}.$$

Typically: $\alpha \approx 1$ and n is very large, and fairly easy to asses. On the other hand p is very small, and is difficult to assess as it requires a lot of information about the genetic make up of a (potentially large) group of people. When n is small it may be possible to test all of the group. A great variety of mistakes have been made in using complex evidence of this type in courts. The famous 'prosecutor's fallacy' is pretending that $P(G^c \mid E) = P(E \mid G^c)$ (of course this is wrong).

We can also combine properties **P4** and **P5** to make a mega-property of conditional expectation: Bayes' theorem for partitions.

Theorem: properties of conditional probability: Bayes theorem for partitions

For any partition A_1, ..., A_k and any B with $\mathbb{P}(B) > 0$,

$$P(A_i \mid B) = \frac{P(A_i)\,P(B \mid A_i)}{\sum_{j=1}^k P(A_j)\,P(B \mid A_j)}.$$

More generally, if $\mathbb{P}(B \mid C) > 0$,

$$P(A_i \mid B \cap C) = \frac{P(A_i \mid C)\,P(B \mid A_i \cap C)}{\sum_{j=1}^k P(A_j \mid C)\,P(B \mid A_j \cap C)}.$$

Try it out

On any given day, it rains with probability 1/2. If it rains, Charlie the cat will go outside with probability 1/10; if it is dry, the probability is 3/5. If Charlie goes outside, what is the conditional probability that it has rained?

Answer:

Let R = it rains, C = Charlie goes outside. Then R, R^c form a partition with $P(R) = P(R^c) = 1/2$. Also, $P(C \mid R) = 1/10$ and $P(C \mid R^c) = 3/5$. So, by Bayes' theorem (P6),

$$\begin{aligned} P(R \mid C) &= \frac{P(C \mid R)\,P(R)}{P(C \mid R)\,P(R) + P(C \mid R^c)\,P(R^c)} \\ &= \frac{\frac{1}{10} \cdot \frac{1}{2}}{\frac{1}{10} \cdot \frac{1}{2} + \frac{3}{5} \cdot \frac{1}{2}} = \frac{1}{7}. \end{aligned}$$

Textbook references

If you want more help with this section, check out:

- Sections 2.3 and 2.4 in (Blitzstein and Hwang 2019);
- Sections 2.1 and 2.2 in (Anderson, Seppäläinen, and Valkó 2018);
- or Section 2.1 in (Stirzaker 2003).

3.3 Independence of events

Tied to the idea of conditional probability is the idea of *independence*: the property that two events are unrelated, or have no bearing on each other's likelihood.

Key idea: Independence of two events

We say that two events A and B are *independent* whenever

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

We say that two events A and B are *conditionally independent* given a third event C with $\mathbb{P}(C) > 0$ whenever

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

For example, if we pick a card from a well-shuffled deck, the events "the card is red" (R) and "the card is an Ace" (A) are independent.

By counting, we have that $P(R) = \frac{26}{52} = \frac{1}{2}$ and $\mathbb{P}(A) = \frac{4}{52} = \frac{1}{13}$. Now, $A \cap R = \{A\diamondsuit, A\heartsuit\}$ so $P(A \cap R) = \frac{2}{52} = \frac{1}{26}$, We check that $\frac{1}{26} = \frac{1}{2} \cdot \frac{1}{13}$, so R and A are indeed independent.

Try it out

Roll two standard dice. Let E be the event that we have an even outcome on the first die. Let F be the event that we have a 4 or 5 on the second die. Are E and F independent?
Answer:
We will verify using a counting argument.
There are 36 equally likely outcomes, namely:

$$\Omega = \{(i, j) \colon i \in \{1, \dots, 6\} \text{ and } j \in \{1, \dots, 6\}\}$$

Of those, 3×6 are in E, and 6×2 are in F, so $P(E) = 18/36 = 1/2$ and $P(F) = 12/36 = 1/3$. Moreover, 3×2 of these outcomes belong to both E and F, so $P(E \cap F) = 6/36 = 1/6$. Indeed,

$$P(E \cap F) = 1/6 = 1/2 \times 1/3 = P(E)\,P(F)\,.$$

So E and F are independent.

Try it out

Roll a fair die. Consider the events

$$A_1 = \{2, 4, 6\},\ A_2 = \{3, 6\},\ A_3 = \{4, 5, 6\},\ \text{and } A_4 = \{1, 2\}.$$

Which pairs of events are independent?

Answer:

Note that $A_1 \cap A_2 = \{6\}$ so $P(A_1 \cap A_2) = \frac{1}{6}$, while $P(A_1)\,P(A_2) = \frac{3}{6} \cdot \frac{2}{6} = \frac{1}{6}$ too. So A_1 and A_2 are independent.

On the other hand, $A_1 \cap A_3 = \{4, 6\}$ so $P(A_1 \cap A_3) = \frac{2}{6} = \frac{1}{3}$, but $P(A_1)\,P(A_3) = \frac{3}{6} \cdot \frac{3}{6} = \frac{1}{4}$. So A_1 and A_3 are *not* independent.

Never confuse disjoint events with independent events! For independent events, we have that $P(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$, but for disjoint events, $P(A \cap B) = 0$ because $A \cap B = \emptyset$.

Disjointness is a property of the *sets* only (it can be seen from the Venn diagram). Independence is a property of *probabilities* (it cannot be seen from the Venn diagram).

Try it out

In the context of the previous example, $A_3 \cap A_4 = \emptyset$, so A_3 and A_4 are disjoint. They are certainly not independent, since $P(A_3 \cap A_4) = 0$ but $P(A_3)\,P(A_4) = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6} \neq 0$.

The next theorem explains why independence is called independence:

Theorem: equivalent forms for independence

Consider any two events A and B with $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$. The following statements are equivalent.

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

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

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

In other words, learning about B will not tell us anything new about A, and similarly, learning about A will not tell us anything new about B.

For conditional independence, we have a similar result.

Theorem: equivalent forms for conditional independence

Consider any three events A, B, and C, with $P(A \cap B \cap C) > 0$. The following statements are equivalent.

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

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

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

In other words, if we know C then learning about B will not tell us anything new about A, and similarly, if we know C then learning about A will not tell us anything new about B.

Consider the card-shuffling example again. The probability that our card is an Ace is $\mathbb{P}(A) = 1/13$ and the probabilitiy that it is an Ace, given it is red, is

$$P(A \mid R) = \frac{P(A \cap R)}{P(R)} = \mathbb{P}(A),$$

by independence. The 'reason' for the independence is that the proportion of aces in the deck (4/52) is the same as that of aces among the red cards (2/26).

Key idea

It is possible for two events to be conditionally independent on particular events, but not to be (unconditionally) independent. We will see an example of this when we discuss genetics, in Section 5.2.

It can be extremely useful to recognize situations where (conditional) independence can be applied. Of course, it is equally important not to assume (conditional) independence where there really are dependencies.

Definition: independence for multiple events

A (possibly infinite) collection of events $\mathcal{A} \subseteq \mathcal{F}$ are *mutually independent* if for every *finite* non-empty $\mathcal{C} \subseteq \mathcal{A}$ (that is, $\mathcal{C}$ is a finite subcollection of the events in question),

$$P\left(\bigcap_{A \in \mathcal{C}} A\right) = \prod_{A \in \mathcal{C}} \mathbb{P}(A).$$

A collection of events $\mathcal{A} \subseteq \mathcal{F}$ are *mutually conditionally independent* given another event B if for every finite non-empty subcollection $\mathcal{C} \subseteq \mathcal{A}$,

$$P\left(\bigcap_{A \in \mathcal{C}} A \mid B\right) = \prod_{A \in \mathcal{C}} \mathbb{P}(A \mid B).$$

The smallest case here is to consider three events. We say that the events A, B, and C are mutually independent if all of the following equalities are satisfied:

$$\begin{aligned} P(A \cap B \cap C) &= \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C), \\ P(A \cap B) &= \mathbb{P}(A)\mathbb{P}(B), \\ P(B \cap C) &= \mathbb{P}(B)\mathbb{P}(C), \\ P(C \cap A) &= \mathbb{P}(C)\mathbb{P}(A). \end{aligned}$$

Suppose we roll 4 dice and their values are independent.

To find the probability that we throw no sixes let A_i be the event 'the ith throw is not a 6'. By assumption A_1, ..., A_4 are independent so

$$P(\text{no sixes on 4 dice}) = P\left(\bigcap_{i=1}^{4} A_i\right) = \prod_{i=1}^{4} P(A_i) = \left(\frac{5}{6}\right)^4.$$

The same result is obtained from the classical model, by selection with replacement.

It is possible for events to be *pairwise* independent without being *mutually* independent, as the next example demonstrates.

Examples: Example

Toss two fair coins. The sample space is $\Omega = \{HH, HT, TH, TT\}$ and each outcome has probability $1/4$.
Let $A = \{HH, HT\}$ be the event that the first coin comes up 'heads', $B = \{HH, TH\}$ the event that the second coin comes up 'heads', and $C = \{HH, TT\}$ the event that the coins come up the same.
Then since $\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C) = 1/2$ and each pairwise intersection has probability $1/4$, it is easy to see that the events are pairwise independent. However, $P(A \cap B \cap C) = P(HH) = 1/4$ which is not the same as $\mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C) = 1/8$, so the three events are *not* mutually independent.
To interpret this in words, if we consider any two of the events, the occurrence of one tells us nothing about the occurrence of the other. As soon as we consider statements involving all three events, however, we see the dependence. For example,

$$P(C \mid A \cap B) = 1,$$

since $A \cap B = \{HH\}$ and $\{HH\} \subseteq C$, compared to the unconditional probability $\mathbb{P}(C) = 1/2$.

Textbook references

If you want more help with this section, check out:

- Section 2.5 in (Blitzstein and Hwang 2019);
- Section 2.3 in (Anderson, Seppäläinen, and Valkó 2018);
- or Section 2.2 in (Stirzaker 2003).

3.4 Historical context

Bayes' theorem is named after the Reverend Thomas Bayes (1701–1761); it was published after his death, in 1763. In our modern approach to probability, the theorem is a very simple consequence of our definitions; however, the result may be interpreted more widely, and is one of the most important results regarding statistical reasoning.

Figure 3.1: Thomas Bayes

4 Interpretations of probability

Goals

1. Understand that there are different ways to interpret probability values.

This chapter covers some different ways in which we can interpret probabilities in the real world. The axioms in Chapter 1 are helpful to determine a framework for mathematical probability, but they leave us lots of room to choose a model within that framework.

We have already discussed one approach in Chapter 2: the "classical" approach, in which each outcome in the sample space is assigned the same probability. This interpretation has some obvious limitations in practice. Often we cannot find a set of outcomes that it is reasonable to think of as a priori equally likely. Therefore, it is essential to have more widely applicable models to deal with uncertainty.

In this chapter, we discuss two more approaches to determining probabilities in real-world applications: the *relative frequencies* approach and the *subjective probability* (or betting) approach. Either can be used, depending on the context, to help us to assign probabilities to events.

The goal of this chapter is to help you to develop more intuition and probabilistic thinking. For the rest of the course, we will assume that we "know" the probability of each event, without worrying too much about how it was determined.

4.1 Relative frequency interpretation

This interpretation applies to **trials** giving chance outcomes of an experiment that can be repeated indefinitely under essentially unchanged conditions and which exhibits **long term regularity**.

Suppose that we run n trials of an experiment with a known list of possible outcomes and the number of trials on which event A occurs is n_A (A is again a set of possible outcomes). The **relative frequency** of occurrence of A is n_A/n.

For example, if we toss a coin 1000 times and observe 490 heads, then the relative frequency of heads is 490/1000.

For some experiments, it may be reasonable to suppose that relative frequencies are stable for very large n.

If we toss a fair coin one billion times, we might expect that the relative frequency of heads after the first few thousand throws would remain very close to 1/2.

As a mathematical idealization, we suppose that there is a unique, empirical limiting value for n_A/n, as n tends to infinity, which we call the *relative frequency probability* of A.

For our coin, the statement $P(\text{heads}) = 1/2$ means 'if we tossed the coin an extremely large number of times, then the proportion of heads would be arbitrarily close to 1/2'.

This interpretation is widely used, especially in physics, where experiments are designed for repeatability and we can expect future trials to behave like those in the past. In this view, probability is a property of the experimental setup and may be "objectively' ' discovered by sufficient repetitions of the experiment.

Amongst the problems with this interpretation are:

- it is often impossible to decide what "essentially unchanged conditions' ' are;
- we often have no way of knowing when limiting frequencies become stable (how many trials should we do to test this?);
- we can only use it in situations that are repeatable.

4.2 Betting interpretation

A very different way of interpreting probability goes by considering probability as a quantification of someone's (yours, mine, your neighbour, …) belief that an event will occur. There are various different ways in which we can measure this belief numerically. Here is one of the simplest.

Your **subjective probability** that A will occur is measured by the amount £p_A that you would consider to be a fair price for the following gamble:

- if A occurs, you receive £1;
- if A does not occur, you receive nothing.

In this interpretation, there are no "true' ' probabilities. Different individuals will have different information relevant to a problem and so may validly make different probability assessments.

For instance, if you say your probability that 'Your Team' wins its next match is 1/2 this means that you view £1/2 as a fair price for the gamble winning you £1 if Your Team wins but otherwise nothing. Others may disagree with you.

Subjective probability ideas are often used by decision makers who have to consider problems concerning unique, non-repeatable events, based on their informed but subjective judgements. The advantages of this interpretation are that probability measures the belief of a subject, and is no longer seen as a property of the experimental setup. Potential issues are that the highest 'buying price' may differ from lowest 'selling price'; a subject may have reason to misrepresent their fair price; placing the bet itself might affect the experiment.

4.3 Interpretation and the axioms

We claimed that the axioms of probability are the same regardless of the interpretation of the probabilities that we are using. **A1** and **A2** are clearly very sensible in any interpretation. The justification of **A3** (and, by extension, **A4**) needs some more thought.

A3 feels intuitive for the classical model of probability by its relation to counting: in the classical model if A contains m_A outcomes and B contains m_B outcomes, with none in common with A, then $m_{A\cup B} = m_A + m_B$. The argument is very similar for the relative frequency model and only slightly more subtle for the betting model.

Textbook references

For more information on these ideas, check out:

- Section 1.2 in (DeGroot and Schervish 2013);
- Chapter 0 in (Stirzaker 2003);
- or (Hájek 2012).

5 Some applications of probability

Goals

1. Understand the meaning of a reliability network.
2. Know how to evaluate the reliability of networks.
3. Understand the probabilistic nature of genetics.
4. Know how to derive the probability of genotypes of children given the genotypes of parents.
5. Know how and when to exploit the general rules of (conditional) probability to solve complex reliability networks and problems in genetics.

5.1 Reliability of networks

Reliability theory concerns mathematical models of systems that are made up of individual components which may be faulty.

If components fail randomly, a key objective of the theory is to determine the probability that the system as a whole works. This will depend on the structure of the system (how the components are organized). This is an important problem in industrial (or other) applications, such as electronic systems, mechanical systems, or networks of roads, railways, telephone lines, and so on.

Once we know how to work out (or estimate) failure probabilities of these systems, we can start to ask more sophisticated questions, such as: How should the system be designed to minimize the failure probability, given certain practical constraints? What is a good inspection, servicing and maintenance policy to maximize the life of the system for a minimal cost?

In this course, to demonstrate an application of the probabilistic ideas we have covered so far, we address the basic question: Given a system made up of finitely many components, what is the probability that the system works? Whether the system functions depends on whether the components function, and on the configuration of those components.

Example

The figure below shows (a) two components in series, (b) three in parallel, (c) a four component system. In each case assume that the system works if it is possible to get from the left end to the right through functioning components.

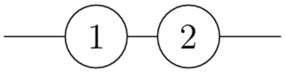


Figure 5.1: a) Two components in series

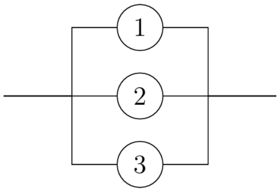


Figure 5.2: b) Three components in parallel

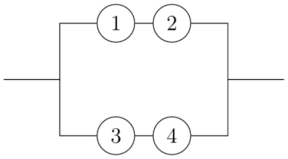


Figure 5.3: c) A system with four components

The system in (a) works if and only if both components 1 and 2 work.
The system in (b) works if any of 1, 2, 3 work.
The system in (c) works if either both 1 and 2 work, or both 3 and 4 work (or all work).

Definition: reliability network

A *reliability network* is a diagram of nodes and arcs. The nodes represent components of a multi-component system, where each node is either working or is broken, and where the entire system works if it is possible to get from the left end to the right of the diagram through working components only. Suppose the ith component functions with probability p_i, $i \in \{1, 2, \dots, k\}$, and different components are independent. The probability that the system works is then a function of the probabilities p_1, ..., p_k. We denote this function by $r(p_1, p_2, \dots, p_k)$, and call it the *reliability function*. It is determined by the layout of the reliability network.

Looking at reliability networks, and determining their reliability functions, is a source of lots of good examples to practice working with the axioms of probability (in particular **A3**, **C1**, and **C6**), as well as building up some more intuition about independence. Let's have one more example (and there are a couple on the problem sheet, too).

Try it out

Consider the three networks in the previous example. We consider the events

$$W_i = \text{component } i \text{ works}, \quad S = \text{system works}.$$

Calculate $P(S)$ for each of the networks.
Suppose that the system in (c) works. What is the conditional probability that component 1 works?
Answer:
The first step is to represent S in terms of the W_i and the operations of set theory. Then we can compute $P(S)$ using our rules for probabilities.
In (a),

$$S = W_1 \cap W_2.$$

By independence, $P(S) = P(W_1)\,P(W_2) = p_1 p_2$.
For (b), we have

$$S = W_1 \cup W_2 \cup W_3.$$

It is easiest to compute

$$P(S^{\mathrm{c}}) = P(W_1^{\mathrm{c}} \cap W_2^{\mathrm{c}} \cap W_3^{\mathrm{c}}) = P(W_1^{\mathrm{c}})\,P(W_2^{\mathrm{c}})\,P(W_3^{\mathrm{c}}) = (1-p_1)(1-p_2)(1-p_3),$$

by independence. So

$$P(S) = 1 - (1-p_1)(1-p_2)(1-p_3).$$

For (c), we have

$$S = (W_1 \cap W_2) \cup (W_3 \cap W_4).$$

Then, by C6,

$$\begin{aligned} P(S) &= P(W_1 \cap W_2) + P(W_3 \cap W_4) - P(W_1 \cap W_2 \cap W_3 \cap W_4) \\ &= p_1p_2 + p_3p_4 - p_1p_2p_3p_4. \end{aligned}$$

To find the conditional probability that component 1 works, given that system (c) works, we go back to the definition of conditional probability:

$$\begin{aligned} P(W_1 \mid S) &= \frac{P(W_1 \cap S)}{P(S)} \\ &= \frac{P((W_1 \cap W_2) \cup (W_1 \cap W_3 \cap W_4))}{P(S)} \\ &= \frac{p_1p_2 + p_1p_3p_4 - p_1p_2p_3p_4}{p_1p_2 + p_3p_4 - p_1p_2p_3p_4}. \end{aligned}$$

Textbook references

If you want more help with this section, check out:

- Sections 4.1–4.4 in (Billinton and Allan 1996);
- or Chapter 9 in (Ross 2010).

5.2 Genetics

Inherited characteristics are determined by *genes*. The mechanism governing inheritance is random and so the laws of probability are crucial to understanding genetics.

Your cells contain 23 pairs of *chromosomes*, each containing many genes (while 23 pairs is specific to humans the idea is similar for all animals and plants). The genes take different forms called *alleles* and this is one reason why people differ (there are also environmental factors). Of the 23 pairs of chromosomes, 22 pairs are *homologous* (each of the pair has an allele for any gene located on this pair). People with different alleles are grouped by visible characteristics into *phenotypes*; often one allele, A say, is *dominant* and another, a, is *recessive* in which case AA and Aa are of the same phenotype while aa is distinct. Sometimes, the recessive gene is rare and the corresponding phenotype is harmful, for example haemophilia or sickle-cell anaemia.

For instance, in certain types of mice, the gene for coat colour (a phenotype) has alleles B (black) or b (brown). B is dominant, so BB or Bb mice are black, while bb mice are brown with no difference between Bb and bB.

With sickle-cell anaemia, allele A produces normal red blood cells but a produces deformed cells. Genotype aa is fatal but Aa provides protection against malarial infection (which is often fatal) and so allele a is common in some areas of high malaria risk.

To apply probability theory to the study of genetics, we use the **basic principle of genetics:** For each gene on a homologous chromosome, a child receives one allele from each parent, where each allele received is chosen independently and at random from each parent's two alleles for that gene.

Examples: Example

For example, in a certain type of flowering pea, flower colour is determined by a gene with alleles R and W, with phenotypes RR (red), RW (pink) and WW (white). The table of offspring genotype probabilities given parental genotypes is

Parental genotype		RR RR	RR RW	RR WW	RW RW	RW WW	WW WW
offspring	RR	1	1/2	0	1/4	0	0
genotype	RW	0	1/2	1	1/2	1/2	0
	WW	0	0	0	1/4	1/2	1

How to read the table: for example, parents RR and RW produce RR offspring with chance 1/2 (the RR parent must supply an R but the RW parent supplies either R or W, each with probability 1/2). When we cross red and white peas, all the offspring will be pink but when we cross red and pink peas, about half of the peas will be red, half pink. Mendel carried out experiments like these to establish the genetic basis of inheritance.

Advanced content

Similar but larger tables are relevant when there are more than two alleles.

It is extremely important to note that **genotypes of siblings are dependent unless we condition on parental genotypes**. For example, if two black mice (which may each be BB or Bb) have 100 black offspring, you may conclude that the next offspring is overwhelmingly likely to also be black, because it is very likely that at least one parent is BB.

Genes can also affect reproductive fitness, as we see in the next example.

Try it out

A gene has alleles A and a but a is recessive and harmful, so genotype aa does not reproduce while AA, Aa are indistinguishable. With proportions $1-\lambda$, λ of AA, Aa in the healthy population, show that the probability of an aa offspring is $\lambda^2/4$.
Answer: To show this we can use the partition F_{AA}, F_{Aa} (father AA, Aa respectively) to calculate

$$P(Fa) = P(F_{AA})\,P(Fa \mid F_{AA}) + P(F_{Aa})\,P(Fa \mid F_{Aa}) = 0 + \lambda \times (1/2) = \lambda/2$$

for the event Fa that the father provides allele a.
By symmetry, the mother also supplies allele a with probability $\lambda/2$ and by independence (random mating) the probability that they both supply allele a is $\lambda^2/4$ e.g. when $\lambda \approx 1/2$, about 6% of

offspring will be aa. Over time the proportion of allele a will decrease unless Aa has a reproductive advantage over AA.

Things are slightly different for genes on the X or Y chromosomes (sex-linked genes).

These are the final chromosome pair, known as the sex chromosomes. Each may be X, a long chromosome, or Y, a short chromosome. Most of the genes on X do not occur on Y. Most people have sex determined as XX (female) or XY (male); YY is not possible.[1]

Try it out

A gene carried on the X chromosome has alleles A and a (so men have only one allele, while women have two).

- aa women are unhealthy;
- a men are unhealthy;
- otherwise the person is healthy.

A male child inherits his gene on the X chromosome from his mother (as he must get his Y from his father) with equal chance of the two alleles that the mother carries. A female child inherits her father's single allele as well as one of her mother's two alleles.
Jane is healthy. Her maternal aunt has an unhealthy son (Jane's cousin). Jane's maternal grandparents and her father are all healthy.

i. What is the probability that Jane is genotype Aa?

Now suppose that Jane has two healthy brothers.

ii. What now is the probability that Jane is genotype Aa?

Answer: We start with part i. From the information given, we can add some information to the genetic tree. The healthy men are A. A male child receives his single (X-carried) allele as a random selection from his mother's two alleles (the genotype of his father has no bearing). Thus Jane's Aunt must carry an a. She cannot have inherited this from the healthy grandfather, so the grandmother must also carry an a. This gives us the picture below.

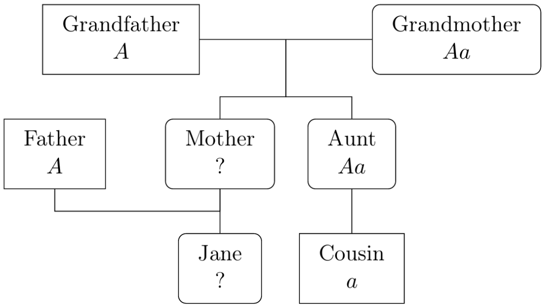


Figure 5.4: Jane's family tree, pt1

[1] At least when restricted to the Riemann integral.

Consider the events $J = \{\text{Jane is } Aa\}$, $M_1 = \{\text{mother is } AA\}$, and $M_2 = \{\text{mother is } Aa\}$. From the tree above, we have M_1 occurs if and only if Jane's mother inherited an A from her mother, i.e., $P(M_1) = 1/2$ and $P(M_2) = 1/2$ too. Given the mother's genotype, we can work out the probabilities for Jane's inheritance. Thus, by the partition theorem,

$$\begin{aligned} P(J) &= P(M_1)\,P(J \mid M_1) + P(M_2)\,P(J \mid M_2) \\ &= \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}. \end{aligned}$$

Now we move on to part ii. The tree is now augmented by the additional information about Jane's siblings:

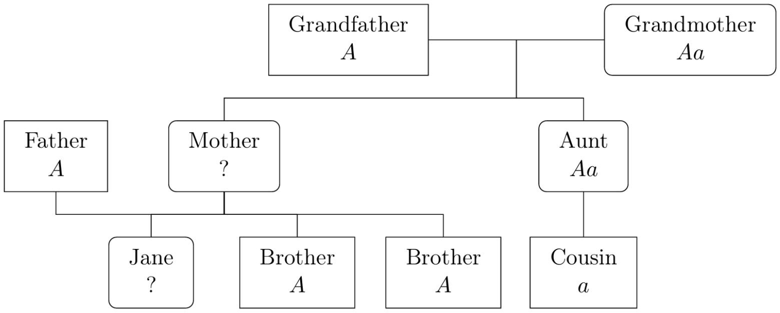


Figure 5.5: Jane's family tree, pt2

Let $B = \{\text{Two brothers are } A\}$. We want $P(J \mid B)$. Note that our knowledge of B changes our beliefs about the genotype of Jane's mother. To see this more clearly, imagine that Jane had 100 brothers, all of whom were of type A. Then we would be very nearly sure that Jane's mother was of type AA, and so Jane would be almost certainly of type AA too.
For the calculation, we use the partition theorem for conditional probabilities:

$$P(J \mid B) = P(M_1 \mid B)\,P(J \mid M_1 \cap B) + P(M_2 \mid B)\,P(J \mid M_2 \cap B)\,.$$

But given M_i, J is independent of B so

$$P(J \mid B) = P(M_1 \mid B)\,P(J \mid M_1) + P(M_2 \mid B)\,P(J \mid M_2)\,.$$

As above, we have $P(J \mid M_1) = 0$ and $P(J \mid M_2) = 1/2$. By Bayes's theorem,

$$\begin{aligned} P(M_2 \mid B) &= \frac{P(B \mid M_2)\,P(M_2)}{P(B \mid M_1)\,P(M_1) + P(B \mid M_2)\,P(M_2)} \\ &= \frac{(1/2)^2 \cdot (1/2)}{1^2(1/2) + (1/2)^2 \cdot (1/2)} = \frac{1}{5}. \end{aligned}$$

So

$$P(J \mid B) = 0 + \frac{1}{2} \cdot \frac{1}{5} = \frac{1}{10}.$$

So seeing that Jane has two healthy brothers significantly reduces the chance that Jane is carrying an a.

Textbook references

If you want more help with this section, check out

- Section V.5 in (Feller 1968);
- or Section 5.6 in (Chung and AitSahlia 2003).

5.3 Hardy-Weinberg equilibrium

Consider a population of a large number of individuals evolving over successive generations. Consider a gene (on a homologous chromosome) with two alleles A and a and genotypes $\{AA, Aa, aa\}$. Suppose the genotype proportions in the population (uniformly for males and females) at generation $n = 0, 1, 2, ...$ are

AA	Aa	aa
u_n	$2v_n$	w_n

where we have $u_n + 2v_n + w_n = 1$. Suppose also that the proportions of the alleles in the population are

A	a
p	q

where $p_n + q_n = 1$. We see that

$$p_n = \frac{2u_n + 2v_n}{2u_n + 4v_n + 2w_n} = u_n + v_n$$

and, similarly, $q_n = v_n + w_n$.

Suppose that

- the gene is *neutral*, meaning that different genotypes have equal reproductive success;
- there is *random mating* with respect to this gene, meaning that each individual in generation $n+1$ draws randomly two parents whose genotypes are independently in the proportions u_n, $2v_n$, w_n.

How do the genotype proportions evolve over successive generations?

Consider the offspring of generation 0. Let FA = event that child gets allele A from father, MA = event that child gets allele A from mother, F_{AA} = event that father is AA, F_{Aa} = event that father is Aa, F_{aa} = event that father is aa. Then

$$\begin{aligned} P(FA) &= P(F_{AA})\,P(FA \mid F_{AA}) + P(F_{Aa})\,P(FA \mid F_{Aa}) + P(F_{aa})\,P(FA \mid F_{aa}) \\ &= 1 \cdot u_0 + \frac{1}{2} \cdot 2v_0 + 0 \cdot w_0 = u_0 + v_0 = p_0. \end{aligned}$$

Similarly, $P(MA) = p_0$. In particular, since parents contribute alleles independently, the probability distribution of the genotype of an individual in generation 1 is

AA	Aa	aa
p_0^2	$2p_0(1-p_0)$	$(1-p_0)^2$

Provided that the population is large enough (see the *law of large numbers* in Chapter 9) these will also be the generation 1 proportions of AA, Aa, aa, i.e.,

$$u_1 = p_0^2, \qquad v_1 = p_0(1-p_0), \qquad w_1 = (1-p_0)^2.$$

Now let $p_1 = u_1 + v_1$ be the proportion of A in the gene pool at generation 1. Substituting the values of u_1, v_1 we find that

$$p_1 = u_1 + v_1 = p_0^2 + p_0(1-p_0) = p_0,$$

i.e. the proportions of A and a in the gene pool are constant.

The same argument applies for later generations, so that $p_n = p_0$ for all n, i.e., the proportions of the two alleles in the gene pool remain constant. This means that, for $n \geq 1$,

$$u_n = p_0^2, \qquad v_n = p_0(1-p_0), \qquad w_n = (1-p_0)^2,$$

so that the proportions of the three genotypes in the population remain constant in every generation after the first. This is called the *Hardy–Weinberg equilibrium.*

5.4 Historical context

Reliability for systems of infinitely many components is related to *percolation.*

On the infinite square lattice $\mathbb{Z}^2$, declare each vertex to be *open*, independently, with probability $p \in [0,1]$, else it is *closed.* Consider the *open cluster* containing the origin, that is, the set of vertices that can be reached by nearest-neighbour steps from the origin using only open vertices. Percolation asks the question: for which values of p is the open cluster containing the origin *infinite* with positive probability? It turns out that for this model, the answer is: for all $p > p_{\mathrm{c}}$ where $p_{\mathrm{c}} \approx 0.593$.

The picture shows part of a percolation configuration, with open sites indicated by black dots and edges between open sites indicated by unbroken lines.

Percolation is an important example of a probability model that displays a *phase transition.* You may see more about it if you do later probability courses.

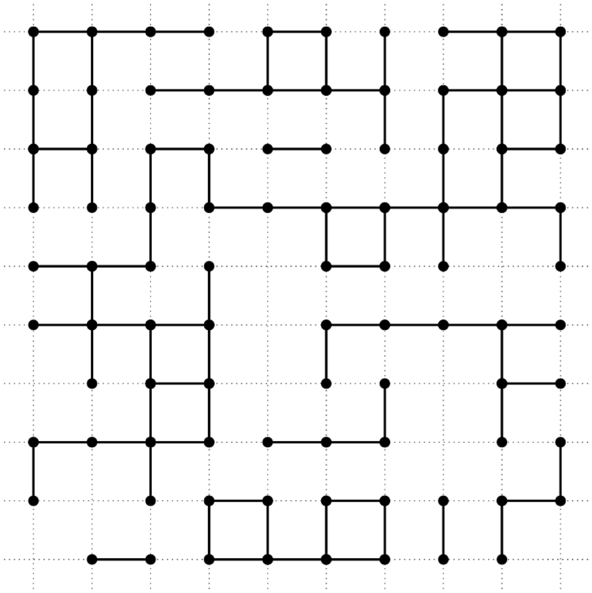


Figure 5.6: A lattice, with some edges missing

The laws governing the statistical nature of inheritance were first observed and formulated by monk Gregor Johann Mendel (1822–1884).

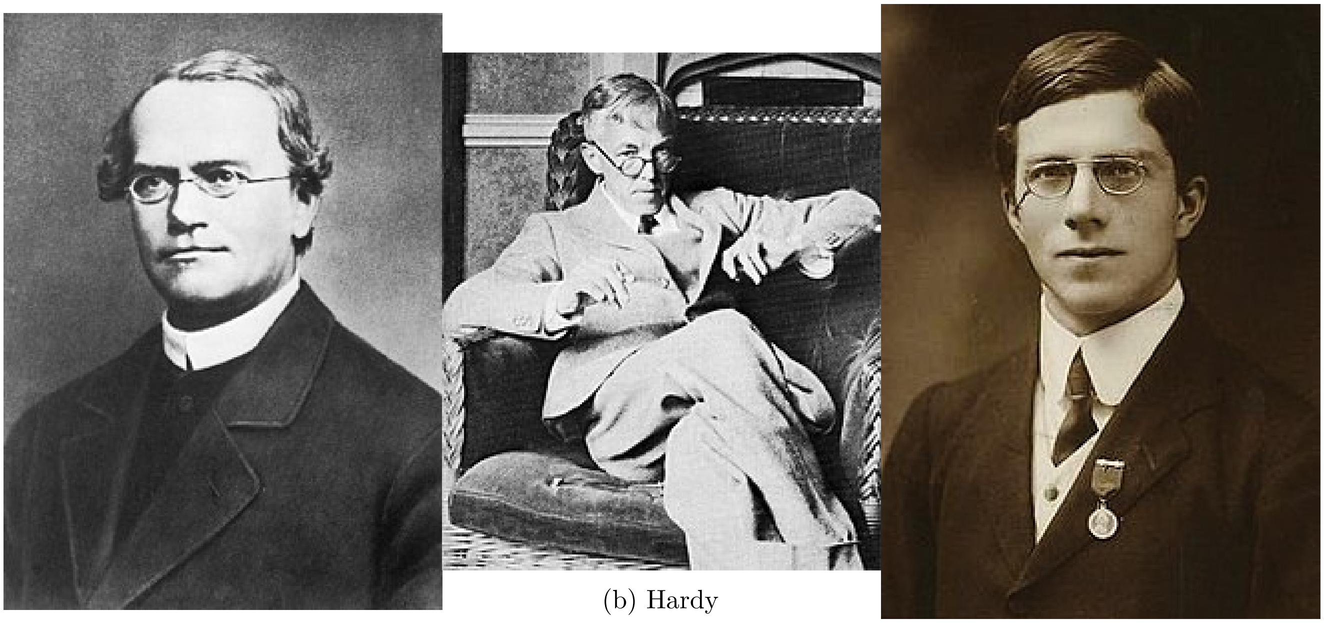
Biologist William Bateson (1861–1926)) coined the terms “genetics” and “allele”.

The Hardy of the Hardy–Weinberg law is G.H. Hardy (1877–1947), the famous mathematical analyst, who published it in 1908.

The statistician R.A. Fisher (1890–1962) made significant contributions to genetics, and much early work in statistics was concerned with genetical problems. A lot of this work contributed to a legacy of eugenics, which was used as a justification for racial discrimination.

The Wright–Fisher model formulates a random model for the evolution of genes in a population with mutation as an *urn model* (Mahmoud 2009, chap. 9).

The deep influence of probability theory on genetics has continued in recent times, with significant developments including the *coalescent* of J.F.C. Kingman.



(a) Mendel

(b) Hardy

(c) Fisher

Figure 5.7: Mendel, Hardy, and Fisher.

6 Random variables

Goals

1. Understand the definition of a random variable as a function on the sample space.
2. Master the notation for events and probabilities of events relating to random variables.
3. Know how to recognize a discrete random variable, how to identify its probability mass function, and how to derive probabilities of associated events.
4. Know how to recognize a continuous real-valued random variable, how to identify its probability density function, and how to derive probabilities of associated events.
5. Know the following distributions. This includes identifying the scenarios in which they hold, the assumptions behind them, and how to identify their parameters.

- the binomial distribution
- the geometric distributions
- the Poisson distribution, including how and when it can be use to approximate a binomial distribution.
- the uniform distribution.
- the exponential distribution, and how it arises from the Poisson distribution.
- the normal distribution, and how we can derive probabilities of events using its cumulative distribution function and standard normal tables.

6. Work with functions of random variables.

In many experiments we are often interested in a numerical value rather than the elementary event $\omega \in \Omega$ *per se*: for example, in the financial industry we may not care about the behaviour of the stock price throughout a given period, only whether it reached a certain level or not; in weather forecasting we may not be interested in the detailed variation of atmospheric pressure and temperature, only in how much rain is going to fall, and so on. These uncertain quantities associated with random scenarios have as their mathematical idealization the concept of *random variable*.

Put simply, a random variable is a *function* or *mapping* of the sample space: for each $\omega \in \Omega$, the random variable X gives the output $X(\omega)$. In this chapter, we study both discrete and continuous univariate random variables, and discuss some important examples: binomial, geometric, Poisson, uniform, exponential, and normal distributions. To enable practical calculations, we also discuss cumulative distribution functions, standard tables, and how probabilities behave under transformations.

6.1 Definition and notation

Key idea: Definition: random variable

A *random variable* on Ω is a mapping from the sample space Ω to some set of *possible values* $X(\Omega) := \{X(\omega) : \omega \in \Omega\}$:

$$X : \Omega \to X(\Omega) \text{ given by } \omega \mapsto X(\omega).$$

Typically, we find ourselves in one of the following situations:

- $X(\Omega) \subseteq \mathbb{R}$, in which case we say that X is a *real-valued random variable* or a *univariate random variable*; or
- $X(\Omega) \subseteq \mathbb{R}^d$, in which case we say that X is a *vector-valued random variable*, or a *multivariate random variable*; in this case we can identify X with a vector $(X_1, \dots, X_d)$ of real-valued random variables, where $X_i(\omega) := [X(\omega)]_i$, that is, $X_i(\omega)$ is the ith component of $X(\omega)$.

For any $B \subseteq X(\Omega)$, we write '$X \in B$' to denote the event $\{\omega \in \Omega : X(\omega) \in B\}$. For any $x \in X(\Omega)$, we write '$X = x$' to mean the event $X \in \{x\}$, that is, $\{\omega \in \Omega : X(\omega) = x\}$. We sometimes also write $\{X = x\}$ and $\{X \in B\}$ to emphasize that these are sets.

For example, if we consider throwing two standard dice, with sample space

$$\Omega = \{(i, j) : i \in \{1, 2, 3, 4, 5, 6\},\ j \in \{1, 2, 3, 4, 5, 6\}\},$$

then the sum of the numbers that show on the dice corresponds to a real-valued random variable X defined by:

$$X(i, j) := i + j, \text{ for all } (i, j) \in \Omega.$$

Then the notation $X = 10$ denotes the event $\{(4, 6), (5, 5), (6, 4)\}$, and $X \in [0, 4]$ denotes the event

$$\{(1, 1), (1, 2), (2, 1), (1, 3), (2, 2), (3, 1)\}.$$

Try it out

Toss 3 fair coins, and let X denote the total number of heads.

1. Describe the function X by tabulating its values:

ω	HHH	HHT	HTH	THH	HTT	THT	TTH	TTT
$X(\omega)$	-	-	-	-	-	-	-	-

2. Consider the events

$$A_1 = \{X = 2\},\ A_2 = \{X \in [0, 1.5]\},\ \text{and } A_3 = \{X \in [10, 20]\}.$$

To which subsets of Ω do these events correspond? What are their probabilities?

Answer:

1. The table of values of X looks like this:

ω	HHH	HHT	HTH	THH	HTT	THT	TTH	TTT
$X(\omega)$	3	2	2	2	1	1	1	0

2. We have that

$$A_1 = \{\omega \in \Omega : X(\omega) = 2\} = \{\text{HHT}, \text{HTH}, \text{THH}\},$$

so, assuming all 8 outcomes are equally likely, $P(X = 2) = 3/8$.

Similarly,

$$A_2 = \{\omega \in \Omega : 0 \le X(\omega) \le 1.5\} = \{\text{TTT}, \text{TTH}, \text{THT}, \text{HTT}\},$$

so $P(X \in [0, 1.5]) = 4/8 = 1/2$, and $A_3 = \emptyset$ so $P(X \in [10, 20]) = 0$.

A simple but important class of random variable is formed by the *indicator random variables*, denoted for an event $A \in \mathcal{F}$ by $\mathbb{1}_A$ and given by the mapping

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

Note that $P(\mathbb{1}_A = 1) = P(\{\omega \in \Omega : \omega \in A\}) = P(A)$ and $P(\mathbb{1}_A = 0) = 1 - P(A)$.

Recall the definition of probability from Section 1.5. Given a probability distribution $P()$ on the sample space Ω, a random variable $X : \Omega \to X(\Omega)$ induces a probability distribution $\mathbb{P}_X(\cdot)$ on the sample space $X(\Omega)$ as follows.

Key idea: Theorem: defining probailities

The function $\mathbb{P}_X(\cdot)$, mapping sets $B \subseteq X(\Omega)$ to a real number $\mathbb{P}_X(B)$, defined by

$$\mathbb{P}_X(B) := P(X \in B) = P(\{\omega \in \Omega : X(\omega) \in B\}),$$

is a probability on $X(\Omega)$, that is, $\mathbb{P}_X(\cdot)$ satisfies the probability axioms (**A1**–**A4**).

Proof

The proof of this theorem is an exercise! It's 6.17 on the problem sheet.

Advanced content

As mentioned in Chapter 1, in general sample spaces Ω one cannot expect to be able to assign probabilities to *every* $A \in 2^\Omega$, and one must restrict to some $\mathcal{F} \subset 2^\Omega$ of 'nice events'.
This has consequences for the definition of random variable we gave above, as can be gleaned by examination of the Theorem: we must have that $\{\omega \in \Omega : X(\omega) \in B\}$ is a genuine event in $\mathcal{F}$, at least for a large enough class of $B \subseteq X(\Omega)$ to tell us everything we would like to know about the random variable X.
For real-valued random variables, the relevant B are the *Borel sets*: these are the members of the smallest σ-algebra on $\mathbb{R}$ that contains all open intervals. Those of you who do Probability II (and later courses) will see that the relevant concept for a more complete definition of random variable is that of a *measurable function*.

Textbook references

If you want more help with this section, check out:

- Section 3.1 in (Blitzstein and Hwang 2019);
- Section 1.5 in (Anderson, Seppäläinen, and Valkó 2018);
- or Section 4.1 in (Stirzaker 2003).

6.2 Discrete random variables

The function $\mathbb{P}_X(\cdot)$ tells us everything we might need to know about the random variable X: it is called the *distribution* of X. In general it is a large and unwieldy object, as we need a value of $\mathbb{P}_X(B)$ for every $B \subseteq X(\Omega)$. However, there are two special cases where we can give an efficient description of the distribution. The first is in the case of *discrete random variables* (the second, which we will see a little later, is the case of *continuous random variables*).

Recall that a set is countable if there exists a bijection between that set and a subset of the natural numbers $\mathbb{N}$.

Key idea: definition: discrete random variable and probability mass function

A random variable $X : \Omega \to X(\Omega)$ is said to be *discrete* when there is a finite or countable set of values $\mathcal{X} \subseteq X(\Omega)$ such that $P(X \in \mathcal{X}) = 1$. The function $p() : \mathcal{X} \to [0, 1]$ defined by

$$p(x) = P(X = x)\,, \text{ for all } x \in \mathcal{X},$$

is called the *probability mass function* of X.

The fact that $p(x) \in [0, 1]$ is an immediate consequence of **A1** and **C4**. Here are some further important properties of the probability mass function.

Key idea: probability mass functions for discrete random variables

Suppose that X is a discrete random variable and $p() : \mathcal{X} \to [0, 1]$ is its probability mass function.

Then

$$P(X \in B) = \sum_{x \in B} p(x), \text{ for all } B \subseteq \mathcal{X},$$

and

$$\sum_{x \in \mathcal{X}} p(x) = 1.$$

Proof

Any $A \subseteq \mathcal{X}$ is finite or countable (because it is a subset of a finite or countable set) and so

$$\{X \in B\} = \bigcup_{x \in B} \{X = x\},$$

where the union runs over a countable number of events. Thus we may apply A4 to get

$$P(X \in B) = \sum_{x \in B} P(X = x) = \sum_{x \in B} p(x).$$

In particular, taking $B = \mathcal{X}$ we get $\sum_{x \in \mathcal{X}} p(x) = P(X \in \mathcal{X}) = 1$.

The probability mass function of a discrete random variable summarizes all information we have about X. Specifically, it allows us to calculate the probability of every event of the form $\{X \in B\}$.

In simple cases, the set $\mathcal{X}$ is often just $X(\Omega)$, but this definition is necessary to cover all cases. If the random variable under consideration is not clear from the context, we may write $p_X()$ for the probability mass function of X.

Theorem: alternative characterisation of discrete random variables

A random variable $X : \Omega \to X(\Omega)$ is discrete whenever

(i) $X(\Omega)$ is finite or countable, or
(ii) Ω is finite or countable.

Proof

Note that (ii) implies (i).
Then, if (i) holds, the statement is immediate from the definitition of a discrete random variable, since one can simply take $\mathcal{X} = X(\Omega)$.

Try it out

Continuing our previous example, toss three fair coins and let X be the total number of heads obtained. This example is discrete since the possible values are 0, 1, 2, 3. Examining the table, and grouping the 8 outcomes by the value of X, we find that the probability mass function of X is

x	0	1	2	3
$p(x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

For example,

$$p(2) = \frac{|\{\text{HHT}, \text{THH}, \text{HTH}\}|}{|\Omega|} = \frac{3}{8}.$$

A quick way to get this is to observe that the number of ways of getting x heads is $\binom{3}{x}$ so $P(X = x) = \binom{3}{x}\frac{1}{8}$ for $x \in X(\Omega) = \{0, 1, 2, 3\}$. We will see shortly that this is an example of the *binomial distribution.*

Textbook references

If you want more help with this section, check out:

- Section 3.2 in (Blitzstein and Hwang 2019);
- Section 3.1 in (Anderson, Seppäläinen, and Valkó 2018);
- or Section 4.2 in (Stirzaker 2003).

6.3 The binomial and geometric distributions

Consider the following random experiment, called a *binomial scenario*:

- A sequence of n trials will be carried out, where n is known in advance of the experiment.
- Trials are independent.
- Each trial has only two outcomes, usually denoted 'success' or 'failure'.
- Each trial succeeds independently with the same probability p.

Consider the random variable X, the total number of successes in the n trials.

The usual sample space Ω for the binomial scenario is the set of all the possible length-n sequences of successes and failures; if we represent a success by 1 and a failure by 0, then each $\omega \in \Omega$ is a string $\omega = \omega_1\omega_2 \cdots \omega_n$ with each $\omega_i \in \{0, 1\}$. The random variable X takes values in $X(\Omega) := \{0, 1, \dots, n\}$, which is finite, so X is a discrete random variable. As a mapping on the sample space, we have $X(\omega) = \sum_{i=1}^{n} \omega_i$, the total numbers of 1s in the string.

For each $x \in \{0, 1, \dots, n\}$:

- because trials are independent (see the definition of independence of multiple events in Section 3.3), every sequence $\omega \in \Omega$ with exactly x successes and $n-x$ failures has probability $P(\{\omega\}) = p^x(1-p)^{n-x}$; and
- there are $\binom{n}{x}$ sequences with exactly x successes.

By **(C7)**, we can sum the probabilities of the outcomes in the event

$$\{X = x\} = \{\omega \in \Omega : \sum_i \omega_i = x\}$$

to obtain

$$p(x) = P(X = x) = \sum_{\omega \in \Omega : \sum_i \omega_i = x} P(\omega)\,.$$

Putting everything together, the probability mass function of X is given by the following, which we take as a definition.

Key idea: Definition: binomial distribution

We say that a discrete random variable X is *binomially distributed* with parameters $n \in \mathbb{N}$ and $p \in [0,1]$, and we write $X \sim \text{Bin}(n, p)$, when $\mathcal{X} = \{0, 1, \dots, n\}$ and

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x} \text{ for all } x \in \{0, 1, 2, \dots, n\}.$$

In the case of just a single trial ($n = 1$), the binomial scenario is often referred to as a *Bernoulli trial*, and $X \sim \text{Bin}(1, p)$ is often referred to as a *Bernoulli random variable* with parameter p.

Example

If we roll 4 fair cubic dice and let X be the number of 6s then $P(X = x) = \binom{4}{x}\left(\frac{1}{6}\right)^x\left(\frac{5}{6}\right)^{4-x}$ so the probability mass function of X is

$$p(0) = \frac{625}{1296}, \quad p(1) = \frac{500}{1296}, \quad p(2) = \frac{150}{1296}, \quad p(3) = \frac{20}{1296}, \quad p(4) = \frac{1}{1296}.$$

Hopefully you can see that these probabilities sum to 1.

Try it out

105 people bought tickets for a flight. Each person independently has chance 0.04 of missing the flight. Find

(a) the probability that nobody misses the flight;

(b) the probability that three or more people miss the flight.

Answer:
Let X be the number of people that miss the flight. Then $X \sim \text{Bin}(105, 0.04)$ so

$$p(x) = \binom{105}{x} \cdot 0.04^x \cdot 0.96^{105-x}.$$

a. We find that $P(X = 0) = p(0) = \binom{105}{0} \cdot 0.04^0 \cdot 0.96^{105} = 0.96^{105} \approx 0.014$.

b. The trick here is to apply **(C2)**, so that

$$P(X \geq 3) = 1 - P(X < 3) = 1 - p(0) - p(1) - p(2),$$

where

$$\begin{aligned} p(0) &\approx 0.014 \text{ as before,} \\ p(1) &= \binom{105}{1} \cdot 0.04^1 \cdot 0.96^{104} \approx 0.060, \\ p(2) &= \binom{105}{2} \cdot 0.04^2 \cdot 0.96^{103} \approx 0.130, \end{aligned}$$

so $P(X \geq 3) \approx 0.796$.

Suppose that we extend the binomial scenario indefinitely, to an unlimited number of trials, and we repeat the trials until we obtain the first success. The (random) number of trials up to and including the first success is called the *geometric distribution.* Note that

$$P(\text{first success occurs on trial } n) = P(\text{first } n-1 \text{ trials are failure, then trail } n \text{ is a success})$$
$$(1-p)^{n-1}p,$$

by independence of the trials and the definition of "independence of multiple events". Note that, provided $p > 0$, this is a probability distribution on $\{1, 2, 3, \dots\}$ because, by the geometric series formula,

$$\sum_{n=1}^{\infty} (1-p)^{n-1}p = p \cdot \frac{1}{1-(1-p)} = 1.$$

Thus we are led to the following definition.

Key idea: Definition: geometric distribution

We say that a discrete random variable X is *geometrically distributed* with parameter $p \in (0, 1]$, and we write $X \sim \text{Geo}(p)$, when $\mathcal{X} = \mathbb{N} := \{1, 2, 3 \dots\}$ and

$$p(x) = (1-p)^{x-1}p, \text{ for all } x \in \{1, 2, 3, \dots\}.$$

Textbook references

If you want more help with this section, check out:

- Section 3.3 in (Blitzstein and Hwang 2019);
- Section 2.4 in (Anderson, Seppäläinen, and Valkó 2018);
- or Section 4.2 in (Stirzaker 2003).

6.4 The Poisson distribution

Key idea: Definition: Poisson distribution

We say that a discrete random variable X is *Poisson distributed* with parameter λ, and we write $X \sim \text{Po}(\lambda)$, when $\mathcal{X} = \mathbb{Z}_+ := \{0, 1, 2, \dots\}$ and

$$p(x) = \frac{e^{-\lambda}\lambda^x}{x!} \quad \text{for } x \in \mathbb{Z}_+.$$

The Poisson distribution is used to model counts of events which occur randomly in time at a constant average rate r per unit time, under some natural assumptions. Specifically,

$$P(\text{event occurs in } (x, x+h)) \approx rh, \text{ for small } h.$$

If X is the count of the number of events over a period of length t then it has distribution $\text{Po}(rt)$. Thus, the interpretation of the parameter λ in $\text{Po}(\lambda)$ is as the *average* number of events: we will return to this more formally later. Typical applications are

- calls to a telephone exchange,

- radioactive decay events,
- jobs at a printer queue,
- accidents at busy traffic intersections,
- earthquakes at a tectonic boundary,
- fish biting at an angler's line.

Try it out

From a particular fleet of aircraft, there have been 32 crashes over a 25-year period. Let W, M, and Y denote the number of crashes in the next week, month, and year, respectively. Assume that a year has 365 days and that a month has 30 days. Suppose that crashes occur at random so that the number of crashes in a particular period can be modelled by a Poisson distribution.

(a) How are W, M, and Y distributed?

(b) Find P(no crashes in the next week).

(c) Find P(no crashes in the next month).

(d) Find P(no crashes in the next year).

Answer:
For part (a), we compute the daily rate of crashes. 25 years is 9125 days. So the daily rate of crashes is $r = \frac{32}{9125} \approx 0.0035$. In a week the average number of crashes is $7 \cdot \frac{32}{9125}$, so $W \sim \text{Po}(\frac{7 \cdot 32}{9125})$. Similarly, $M \sim \text{Po}(\frac{30 \cdot 32}{9125})$ and $Y \sim \text{Po}(\frac{365 \cdot 32}{9125})$.
For part (b), the probability that a $\text{Po}(\lambda)$ random variable takes value 0 is $e^{-\lambda}$. So $P(W = 0) = e^{-\frac{7 \cdot 32}{9125}} \approx 0.976$.
Similarly, for part (c) we get $P(M = 0) = e^{-\frac{30 \cdot 32}{9125}} \approx 0.900$, and for part (d) we get $P(Y = 0) = e^{-\frac{365 \cdot 32}{9125}} \approx 0.278$.

Another situation where the Poisson distribution arises is as an approximation to the binomial distribution when p is small and n is large, i.e., events are rare. More precisely, we have the following result.

Key idea: Theorem: Poisson approximation for Binomial distributions

Consider any $\lambda > 0$. Let $X_n \sim \text{Bin}(n, p_n)$ where $\lim_{n \to \infty} np_n = \lambda$, and let $Y \sim \text{Po}(\lambda)$. Then for all $x \in \mathbb{Z}_+$,

$$\lim_{n \to \infty} p_{X_n}(x) = p_Y(x).$$

We describe this by saying that X_n *converges in distribution* to Y.

Proof

Note that since $np_n \to \lambda$, we have $p_n \to 0$. For fixed x we have that, for $n \geq x$,

$$P(X_n = x) = \binom{n}{x} (p_n)^x (1 - p_n)^{n-x} = n^{-x} \frac{n!}{(n-x)!x!} (np_n)^x (1 - p_n)^{n-x},$$

where we observe that, by some calculus of limits,

$$\begin{aligned}\lim_{n\to\infty}(np_n)^x &= \lambda^x,\\ \lim_{n\to\infty} n^{-x}\frac{n!}{(n-x)!} &= \lim_{n\to\infty}\frac{n}{n}\cdot\frac{n-1}{n}\cdots\frac{n-x+1}{n} = 1,\\ \lim_{n\to\infty}(1-p_n)^n &= \lim_{n\to\infty}\left(1-\frac{np_n}{n}\right)^n = e^{-\lambda},\\ \lim_{n\to\infty}(1-p_n)^{-x} &= 1.\end{aligned}$$

Collecting up terms gives

$$\lim_{n\to\infty} P(X_n = x) = \frac{e^{-\lambda}\lambda^x}{x!},$$

as claimed.

This means that if $X \sim \text{Bin}(n, p)$, where n is large and p is small, then *approximately* $X \sim \text{Po}(np)$. As a rule of thumb, with $p \leq 0.05$, we find $n = 20$ gives a reasonable approximation, while $n = 100$ gives a good approximation. The approximation is useful when it allows us to sidestep calculating $\binom{n}{x}$ for large values of n and x.

Example

A typist produces a page of 1000 symbols but has probability 0.001 of mistyping any single symbol and such errors are independent (note: neither assumption is particularly realistic). The probability that a page contains more than two mistakes is $P(X > 2)$ where X is the number of mistakes on the page.
We can use a Binomial distribution: $X \sim \text{Bin}(1000, 0.001)$. As n is large and p is small, we approximately have that $X \sim \text{Po}(1000 \times 0.001) = \text{Po}(1)$. Therefore,

$$P(X > 2) = 1 - P(X \leq 2) \approx 1 - e^{-1}\left(\frac{1^0}{0!} + \frac{1^1}{1!} + \frac{1^2}{2!}\right) \approx 0.0803.$$

Try it out

From 1979 to 1981, 1103 Bristol postmen reported 245 dog-biting incidents (the dogs biting the postmen, that is). In all 191 postmen were bitten, 145 of them just once. Are these numbers consistent with dogs attacking postmen at random?
Answer:
Let $X =$ number of incidents suffered by a particular postman. Supposing that dogs attack at random, then each of the 245 incidents is a random trial, where our postman has chance $1/1103$ to be involved. So $X \sim \text{Bin}(245, 1/1103)$. This is suitable for a Poisson approximation with $\lambda = np = \frac{245}{1103} \approx 0.222$. Approximately, $X \sim \text{Po}(0.222)$, and then

$$\begin{aligned}&P(X = 0) \approx e^{-0.222} \approx 0.80,\\ &P(X = 1) \approx 0.222e^{-0.222} \approx 0.18,\\ &P(X \geq 2) \approx 1 - 0.80 - 0.18 = 0.02.\end{aligned}$$

Compare this to the observed data for the proportion of postmen who were

$$\begin{aligned}\text{not attacked: } & \frac{1103-191}{1103} \approx 0.83,\\ \text{once attacked: } & \frac{145}{1103} \approx 0.13,\\ \text{more than once attacked: } & \frac{191-145}{1103} \approx 0.04.\end{aligned}$$

This looks like a reasonably good fit. One can test this using a *goodness of fit test* that you may have seen in Statistics courses.

Textbook references

If you want more help with this section, check out:

- Section 4.7 in (Blitzstein and Hwang 2019);
- Section 4.4 in (Anderson, Seppäläinen, and Valkó 2018);
- or Section 4.2 in (Stirzaker 2003).

6.5 Continuous random variables

Key idea: Definition: continuous random variables

Consider a real-valued random variable $X : \Omega \to \mathbb{R}$. We say that X is a *continuous random variable*, or that X has a *continuous probability distribution*, or that X is *continuously distributed*, when there is a non-negative function $f : \mathbb{R} \to \mathbb{R}$ such that

$$P(X \in [a,b]) = \int_a^b f(t)\,\mathrm{d}t, \tag{6.1}$$

for all $[a,b] \subseteq \mathbb{R}$. In this case, f is called the *probability density function* of X.

If X is continuously distributed, then taking $a = b = x$ in Equation 6.1 we see that $P(X = x) = 0$ for all $x \in \mathbb{R}$, and so for any $a < b$ we have

$$P(X \in [a,b]) = P(X \in [a,b)) = P(X \in (a,b]) = P(X \in (a,b))\,.$$

Roughly speaking, $f(x)$ has the interpretation

$$P(X \in [x, x + \mathrm{d}x]) = f(x)\,\mathrm{d}x, \text{ for all } x \text{ at which } f \text{ is continuous.} \tag{6.2}$$

All of the continuous random variables that we will see in this course have a probability density function that is *piecewise continuous*.

If the random variable under consideration is not clear from the context, we may write $f_X(\cdot)$ for the probability density function of X.

Advanced content

A probability density function satisfying Equation 6.1 is not quite unique. In fact, we can change the value of $f(x)$ at countably many x and not affect Equation 6.1. For example, the two probability density functions

$$f_1(x) = \begin{cases} 1 & \text{if } 0 \le x \le 1, \\ 0 & \text{otherwise,} \end{cases}, \text{ and } f_2(x) = \begin{cases} 1 & \text{if } 0 < x < 1, \\ 0 & \text{otherwise,} \end{cases},$$

both integrate to 1 and both give rise to the same distribution, i.e., the same collection of $P(X \in B)$ over sensible sets B.

The probability density function of a random variable determines its probability distribution for all events in a reasonable collection of subsets of $\mathbb{R}$ (but not *all* subsets of $\mathbb{R}$). The relevant concept here is again that of a σ-algebra. We will not give the exact definition of this σ-algebra here, but events of the following type can be assigned probabilities.

Key idea: Theorem: density functions determine probabilities

If X is continuously distributed with probability density function $f(\cdot)$, then for any $B \subseteq \mathbb{R}$ that is a finite union of intervals,

$$P(X \in B) = \int_B f(x)\,\mathrm{d}x.$$

The probability density function of a continuous random variable summarizes practically all information we have about X. Specifically, it allows us to calculate the probability of every event of the form $\{X \in B\}$ where B is a finite union of intervals.

Advanced content

The conclusion of this theorem can be extended to all B in the *Borel σ-algebra*, which is the smallest σ-algebra on $\mathbb{R}$ that contains all intervals. Pretty much every subset of $\mathbb{R}$ that you will ever encounter is Borel.

Note that we can also evaluate probabilities of unbounded intervals:

$$\begin{aligned} P(-\infty < X \le b) &= \int_{-\infty}^{b} f(x)\,\mathrm{d}x; \\ P(a \le X < \infty) &= \int_{a}^{\infty} f(x)\,\mathrm{d}x; \\ P(-\infty < X < \infty) &= \int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1. \end{aligned}$$

To see this, we can for example use **(C9)** (continuity along monotone limits) to get

$$\begin{aligned} P(a \le X < \infty) &= P(\cup_{n=1}^{\infty}\{a \le X \le a+n\}) \\ &= \lim_{n\to\infty} P(a \le X \le a+n) \\ &= \lim_{n\to\infty} \int_{a}^{a+n} f(x)\,\mathrm{d}x \\ &= \int_{a}^{\infty} f(x)\,\mathrm{d}x, \end{aligned}$$

as claimed. In particular, we recover a version of **(C10)** for probability density functions:

Theorem: Corollary: densities integrate to one

Let X be a continuous random variable. Then its probability density function $f(\cdot)$ integrates to one:

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1.$$

Try it out

Suppose that a continuous random variable X has probability density given by $f(x) = kx$ for $x \in [0,2]$, where k is some constant, and $f(x) = 0$ for $x \notin [0,2]$. Find k and then compute $P(X \in [0,1])$.
Answer:
To find k, we use the fact that the density integrates to one:

$$1 = \int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = \int_0^2 kx\,\mathrm{d}x = 2k,$$

and therefore $k = 1/2$. With $B = [0,1]$, we have $P(X \in B) = \int_0^1 x/2\,\mathrm{d}x = 1/4$.

Try it out

Suppose that a continuous random variable X has probability density given by

$$f(x) = \begin{cases} k(1+x) & \text{if } -1 \le x < 0, \\ k(2-x) & \text{if } 0 \le x \le 2, \\ 0 & \text{elsewhere,} \end{cases}$$

where k is some constant. Find k and then compute $P(X \in [0,1])$.
Answer:
To find k, note that

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} f(x)\,\mathrm{d}x \\ &= \int_{-1}^{0} k(1+x)\,\mathrm{d}x + \int_0^2 k(2-x)\,\mathrm{d}x \\ &= k\left[x + \frac{x^2}{2}\right]_{-1}^{0} + k\left[2x - \frac{x^2}{2}\right]_0^2 \\ &= \frac{k}{2} + 2k = \frac{5k}{2}, \end{aligned}$$

so $k = 2/5$. Then

$$P(X \in [0,1]) = \int_0^1 \frac{2}{5}(2-x)\,\mathrm{d}x = \frac{2}{5}\left[2x - \frac{x^2}{2}\right]_0^1 = \frac{3}{5}.$$

There are random variables that are *neither* discrete nor continuous, but they do not arise in practice very often. Can you think how to construct one? We will return to this briefly in Section 6.9.

Textbook references

If you want more help with this section, check out:

- Section 5.1 in (Blitzstein and Hwang 2019);
- Section 3.1 in (Anderson, Seppäläinen, and Valkó 2018);
- or Section 7.1 in (Stirzaker 2003).

6.6 The uniform distribution

Key idea: Definition: uniform distribution

Let a and b be real numbers with $a < b$. We say a continuous random variable X is *uniformly distributed* on $[a, b]$, and we write $X \sim \mathrm{U}(a, b)$, when

$$f(x) = \begin{cases} 1/(b-a) & \text{for all } x \in [a, b], \\ 0 & \text{elsewhere.} \end{cases}$$

When $X \sim \mathrm{U}(a, b)$ then X can take any value in the continuous range of values from a to b and the probability of finding X in any interval $[x, x + h] \subseteq [a, b]$ does not depend on x.

Try it out

Suppose that $X \sim \mathrm{U}(0, 3)$. What is $P(X \leq 1)$?
Answer:
We compute

$$P(X \leq 1) = \int_{-\infty}^{1} f(x)\,\mathrm{d}x = \int_{0}^{1} \frac{1}{3}\,\mathrm{d}x = \frac{1}{3}.$$

Textbook references

If you want more help with this section, check out:

- Section 5.2 in (Blitzstein and Hwang 2019);
- Section 3.1 in (Anderson, Seppäläinen, and Valkó 2018);
- or Section 7.1 in (Stirzaker 2003).

6.7 The exponential distribution

Suppose a bell chimes randomly in time at rate $\beta > 0$. Let $T > 0$ denote the time of the first chime (a random variable). The events $\{\text{no chimes in } [0, \tau]\}$ and $\{T > \tau\}$ are the same. We know that the number of chimes in the interval $[0, \tau]$ is $\mathrm{Po}(\beta\tau)$ and so the probability of no chimes is $e^{-\beta\tau}$. Hence

$$P(T > \tau) = e^{-\beta\tau} \quad \text{or} \quad P(T \leq \tau) = 1 - e^{-\beta\tau} \quad \text{for all } \tau \geq 0$$

As $1 - e^{-\beta\tau} = \int_0^\tau \beta e^{-\beta t}\,\mathrm{d}t$ we see that T is a continuous random variable with probability density function $f(t) = \beta e^{-\beta t}$ for all $t \geq 0$.

Key idea: Definition: exponential distribution

Let $\beta > 0$. We say a continuous random variable X is *exponentially distributed* with parameter β, and we write $X \sim \mathcal{E}(\beta)$, when

$$f(x) = \begin{cases} \beta e^{-\beta x} & \text{for all } x \geq 0, \\ 0 & \text{elsewhere.} \end{cases}$$

Textbook references

If you want more help with this section, check out:

- Section 5.5 in (Blitzstein and Hwang 2019);
- Section 4.5 in (Anderson, Seppäläinen, and Valkó 2018);
- or Section 7.1 in (Stirzaker 2003).

6.8 The normal distribution

The normal distribution is one of the most important probability distributions for several reasons, some of which we will see later in this course. As fate would have it, its density is also a little more complicated.

Key idea: Definition: normal distribution

Let μ, σ be real numbers with $\sigma > 0$. We say a continuous random variable X is *normally distributed* with parameters μ and σ^2, and we write $X \sim \mathcal{N}(\mu, \sigma^2)$, when

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \text{ for all } x \in \mathbb{R}.$$

The normal distribution is also sometimes called the *Gaussian* distribution. The probability density function of the normal distribution is bell shaped. Roughly speaking, the parameter μ determines the *location* of the bell, and σ determines the *spread* of the bell: for smaller σ, the bell is narrower. The relevant formal concepts are expectation and standard deviation, which will be introduced later on.

Unfortunately, there is no closed analytical form for $P(X \in [a, b]) = \int_a^b f(x)\,\mathrm{d}x$ when X is normally distributed.

We can compute $P(X \in [a, b])$ by numerical integration, but we would like a quick reference to these computations. This looks unwieldy, since there are four parameters upon which $P(X \in [a, b])$ depends: a, b, μ, and σ. We show by a series of steps how to reduce everything to a *single* parameter, so that values can be easily tabulated. The relevant concepts that we will need to carry out this simplification are the cumulative distribution function, and transformations of random variables. Both of these concepts are important for many other applications too, so we spend a little time on them.

Textbook references

If you want more help with this section, check out:

- Section 5.4 in (Blitzstein and Hwang 2019);
- Section 3.5 in (Anderson, Seppäläinen, and Valkó 2018);

- or Section 7.1 in (Stirzaker 2003).

6.9 Cumulative distribution functions

Key idea: Definition: cumulative distribution function

For any real-valued random variable X, the function $F\colon \mathbb{R} \to [0,1]$ defined by

$$F(x) := P(X \leq x) \text{ for } x \in \mathbb{R}$$

is called the *cumulative distribution function* of X.

If the random variable under consideration is not clear from the context, we may write F_X for the cumulative distribution function of X.

By **(C1)**, it follows that

$$P(X \in (a,b]) = P(X \leq b) - P(X \leq a) = F(b) - F(a).$$

Now for a continuous random variable X this is the same as $P(X \in (a,b))$, $P(X \in [a,b])$, and so on, so knowledge of the cumulative distribution function is sufficient to calculate the probability of virtually any event of practical interest, that is, any finite union of intervals. This is also true for the discrete case, but is a little more complicated as it requires a limiting procedure. The continuous case is thus somewhat simpler, so we start with that.

Key idea: Theorem: cumulative distribution functions and probability density functions

Suppose that X is a continuously distributed random variable on $\mathbb{R}$ with probability density function $f(\cdot)$. Then $F(\cdot)$ is a continuous function and, for all $x \in \mathbb{R}$,

$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t, \quad f(x) = \frac{\mathrm{d}F}{\mathrm{d}x}(x) \text{ when } f(\cdot) \text{ is continuous at } x. \tag{6.3}$$

Proof

The first equality follows from the definition of continuous random variables, and it implies that F is continuous. Then the second equality is a consequence of the fundamental theorem of calculus (correspondence between derivative and integral).

Try it out

Recall from the final example in Section 6.5 that we have a continuous random variable X with a piecewise continuous density given by

$$f(x) = \begin{cases} \frac{2}{5}(1+x) & \text{if } -1 \leq x \leq 0, \\ \frac{2}{5}(2-x) & \text{if } 0 < x \leq 2, \\ 0 & \text{elsewhere.} \end{cases}$$

Find $F(x)$ for all $x \in \mathbb{R}$, and use it to calculate $P(0 \leq X \leq 1)$.

Answer:
Since $f(\cdot)$ is defined piecewise, we must compute F piecewise too, always using $F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$. The sensible way to do this is to work 'left to right', since then we can use the fact that for $x > x_0$,

$$F(x) = \int_{-\infty}^{x_0} f(t)\,\mathrm{d}t + \int_{x_0}^{x} f(t)\,\mathrm{d}t = F(x_0) + \int_{x_0}^{x} f(t)\,\mathrm{d}t,$$

where we choose x_0 to correspond to the piecewise definition of $f(\cdot)$. To start with, for $x \leq -1$, $F(x) = \int_{-\infty}^{x} 0\,\mathrm{d}t = 0$. Next suppose that $-1 \leq x \leq 0$. Then

$$\begin{aligned} F(x) &= F(-1) + \int_{-1}^{x} \frac{2}{5}(1+t)\,\mathrm{d}t \\ &= 0 + \frac{2}{5}\left[t + \frac{t^2}{2}\right]_{-1}^{x} = \frac{1}{5}(x+1)^2. \end{aligned}$$

Now suppose that $0 \leq x \leq 2$. Then

$$\begin{aligned} F(x) &= F(0) + \int_{0}^{x} \frac{2}{5}(2-t)\,\mathrm{d}t \\ &= \frac{1}{5} + \frac{2}{5}\left[2t - \frac{t^2}{2}\right]_{0}^{x} \\ &= 1 - \frac{(x-2)^2}{5}. \end{aligned}$$

Finally, if $x \geq 2$,

$$F(x) = F(2) + \int_{2}^{x} 0\,\mathrm{d}t = F(2) = 1.$$

So we conclude that

$$F(x) = \begin{cases} 0 & \text{if } x \leq -1, \\ \frac{(x+1)^2}{5} & \text{if } -1 \leq x \leq 0, \\ 1 - \frac{(x-2)^2}{5} & \text{if } 0 \leq x \leq 2, \\ 1 & \text{if } x \geq 2. \end{cases}$$

Note that, as expected, F is continuous everywhere and differentiable at all but finitely many points. Now, since X is continuous,

$$\begin{aligned} P(0 \leq X \leq 1) &= P(0 < X \leq 1) \\ &= P(X \leq 1) - P(X \leq 0) \\ &= F(1) - F(0) \\ &= \frac{4}{5} - \frac{1}{5} = \frac{3}{5}. \end{aligned}$$

Try it out

Recall from the first example in Section Section 6.5 that the continuous random variable X has $f(x) = x/2$ for $x \in [0,2]$ and $f(x) = 0$ elsewhere. By the same method as the previous example, the cumulative distribution function is

$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t = \begin{cases} 0 & \text{if } x < 0, \\ x^2/4 & \text{if } x \in [0,2], \\ 1 & \text{if } x > 2. \end{cases}$$

Now let us move on to the case of a discrete random variable. Let us start with an example.

Example

Suppose X is a discrete random variable taking values in $\{0,1,4\}$ with $p(0) = \frac{1}{2}$, $p(1) = p(4) = \frac{1}{4}$. Now $F(x) = P(X \leq x)$ so that $F(x) = 0$ for $x < 0$. At $x = 0$, we have $F(0) = P(X \leq 0) = p(0) = \frac{1}{2}$, so the cumulative distribution function *jumps* and the magnitude of the jump is the value of the probability mass function at that point. This is the general picture. Then $F(x) = \frac{1}{2}$ for all $x \in [0,1)$, until $F(1) = p(0) + p(1) = \frac{3}{4}$. Continuing, we get

$$F(x) = \begin{cases} 0 & \text{if } x < 0, \\ \frac{1}{2} & \text{if } 0 \leq x < 1, \\ \frac{3}{4} & \text{if } 1 \leq x < 4, \\ 1 & \text{if } x \geq 4. \end{cases}$$

Key idea: Theorem: properties of discrete cdfs

If X is a discrete real-valued random variable with probability mass function $p()$, then F is piecewise constant, and

$$F(x) = \sum_{t\colon t \leq x} p(t) \quad p(x) = F(x) - F(x^-).$$

Here $F(x^-)$ means the *limit from the left* $F(x^-) = \lim_{y \uparrow x} F(y)$.

Proof

The first equality follows from the definition of discrete random variables.. Now

$$p(x) = P(X = x) = P(\{X \leq x\} \setminus \{X < x\}) = F(x) - P(X < x)\,.$$

Let x_n be an increasing sequence with $x_n < x$ and $x_n \to x$. Then to prove the theorem it is enough to show that

$$P(X < x) = F(x^-) = \lim_{n\to\infty} F(x_n).$$

Note that since $x_{n+1} > x_n$, $F(x_n) \leq F(x_{n+1}) \leq 1$, so $F(x_n)$ is bounded and increasing, so the limit does exist. Moreover, $X < x$ if and only if $X \leq x_n$ for some n in $\mathbb{N}$, i.e.,

$$P(X < x) = P(\cup_{n=1}^{\infty} \{X \leq x_n\}) = \lim_{n\to\infty} P(X \leq x_n)\,,$$

by C9 (continuity along monotone limits), because the events $A_n = \{X \leq x_n\}$ are increasing in n. But $P(X \leq x_n) = F(x_n)$ and we are done.

We state a result on the properties of the cumulative distribution function in general, that are already apparent from our examples in the special cases of discrete and continuous random variables.

Theorem: properties of continuous cdfs

Let F be the cumulative distribution function of a real-valued random variable. Then F has the following properties.
F1: $\lim_{t\to-\infty} F(t) = 0$ and $\lim_{t\to+\infty} F(t) = 1$.
F2: *Monotonicity.* For any $s \le t$, $F(s) \le F(t)$.
F3: *Right-continuity.* For any $t \in \mathbb{R}$, $F(t) = F(t^+)$ where $F(t^+)$ is the limit *from the right* $F(t^+) = \lim_{s\downarrow t} F(s)$.

We do not give the proof here, but you can try to prove this or consult the recommended text books.

We have already seen, in the discrete and continuous cases, that we can recover the probability mass function and probability density function, respectively, from the cumulative distribution function. In other words, the cumulative distribution function *determines the distribution*. This is true in general.

Theorem: cdfs determine distributions

The cumulative distribution function F of a real-valued random variable X completely determines the distribution of X.

Advanced content

In general this means that F determines $P(X \in B)$ for all Borel sets B.

Now we can see that there are some cumulative distribution functions that correspond to random variables that are neither discrete nor continuous. For example, we might have some jumps but also some continuously increasing parts. There are even examples where the cumulative distribution function is continuous but no probability density function exists: these *singular* distributions are quite pathological and rarely occur in practice.

Textbook references

If you want more help with this section, check out:

- Section 3.6 in (Blitzstein and Hwang 2019);
- Section 3.2 in (Anderson, Seppäläinen, and Valkó 2018);
- or Section 7.2 in (Stirzaker 2003).

6.10 Standard normal tables

Definition: standard normal distribution

A continuous random variable Z is *standard normally distributed* when $Z \sim \mathcal{N}(0,1)$.
In other words, a standard normal is normal with $\mu = 0$ and $\sigma = 1$.

Because the standard normal distribution plays such a central role in many practical probability calculations, we use a special symbol to denote its probability density function and cumulative distribution function:[1]

$$\phi(z) := f(z) = \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2}, \quad \Phi(z) := F(z) = \int_{-\infty}^{z} \phi(t)\, \mathrm{d}t.$$

As already mentioned, Φ has no closed analytical form, however, it can be tabulated. Such tabulation is called a *standard normal table*. Because $\phi(z) = \phi(-z)$, it follows that $\Phi(z) = 1 - \Phi(-z)$, and so we only need to tabulate Φ for non-negative values of z. Some values that are often useful are:

z	0	1.28	1.64	1.96	2.58
$\Phi(z)$	0.5	0.9	0.95	0.975	0.995

Try it out

Suppose $Z \sim \mathcal{N}(0, 1)$. Calculate $P(-1.28 \le Z \le 1.64)$.
Answer:
We compute, making good use of symmetry,

$$\begin{aligned} P(-1.28 \le Z \le 1.64) &= P(Z \le 1.64) - P(Z \le -1.28) \\ &= P(Z \le 1.64) - P(Z \ge 1.28) \\ &= P(Z \le 1.64) - (1 - P(Z \le 1.28)) \\ &= \Phi(1.64) - (1 - \Phi(1.28)) \\ &= 0.95 - (1 - 0.9) = 0.85. \end{aligned}$$

We can use normal tables for Φ to also calculate $P(X \in [a, b])$ for *any* $X \sim \mathcal{N}(\mu, \sigma^2)$. To explain this, we need to look at transformations of random variables, introduced next.

Textbook references

If you want more help with this section, check out:

- Section 5.4 in (Blitzstein and Hwang 2019);
- or Section 3.5 in (Anderson, Seppäläinen, and Valkó 2018).

6.11 Functions of random variables

Suppose $X\colon \Omega \to X(\Omega)$ is a random variable, and $g\colon X(\Omega) \to \mathcal{S}$ is some function. Then $g(X)$ is also a random variable, namely the outcome to a 'new experiment' obtained by running the 'old experiment' to produce a value x for X, and then evaluating $g(x)$. Formally, as a function of $\omega \in \Omega$, $g(X) := g \circ X$, i.e.,

$$g(X)(\omega) := g(X(\omega)) \text{ for all } \omega \in \Omega.$$

For example,

$$P(g(X) \in B) = P(\{\omega \in \Omega\colon g(X(\omega)) \in B\}) \text{ for all } B \subseteq \mathcal{S}.$$

[1] At least when restricted to the Riemann integral.

Examples

1. For any random variable X, we can consider $\sin(X)$, e^{3X}, X^3, and so on, which are all again random variables.

2. Let X be the score when you roll a fair die and let $Y = (X-3)^2$. Then Y is a discrete random variable with probability mass function

y	0	1	4	9
$p(y)$	1/6	1/3	1/3	1/6

and zero elsewhere. To see this, note that $\{Y = 4\} = \{X \in \{1, 5\}\}$, and so on.

3. If X is $\text{Bin}(n, p)$, then $n - X$ is $\text{Bin}(n, 1-p)$, as shown in Exercise 6.6.

4. Let $X \sim \text{U}(0, 1)$. For any constants a and $b > 0$, define $Y := a + bX$. Then $Y \sim \text{U}(a, a+b)$, because for any $x \in [0, 1]$,

$$\{Y \leq a + bx\} = \{a + bX \leq a + bx\} = \{X \leq x\}$$

so that $P(Y \leq a + bx) = P(X \leq x) = x$, and consequently $F(y) = P(Y \leq y) = \frac{y-a}{b}$ whenever $y \in [a, a+b]$. Therefore, by Equation 6.3, indeed, $f(y) = 1/b$ for $y \in [a, a+b]$ (and zero elsewhere), so Y is uniformly distributed on $[a, a+b]$ by the definition of the Uniform distribution.

A similar result holds when $b < 0$.

Example

If $U \sim \text{U}(0, 1)$ and X is a continuous random variable with a cumulative distribution function F_X which is strictly increasing on $[a, b]$, with $F + X(a) = 0$ and $F_X(b) = 1$, then F_X^{-1} exists and $F_X^{-1}(U)$ has the same distribution as X because for any $x \in [a, b]$,

$$P(F_X^{-1}(U) \leq x) = P(U \leq F_X(x)) = F_X(x) = P(X \leq x)\,.$$

This is a special case of the *probability integral transform* and is very useful for generating random samples of X with computer generated 'uniform random numbers'. For example, $-\frac{1}{\beta}\log(1-U)$ is $\mathcal{E}(\beta)$ and so is $\frac{1}{\beta}\log 1/U$ (as U and $1-U$ are both $\text{U}(0, 1)$).

There is one particularly important function which enables us to get the cumulative distribution function of any normally distributed random variable, using just the the standard normal tables (i.e. Φ, the cumulative distribution function of the standard normal).

Key idea: Theorem: standardizing the normal distribution

Suppose $\mu \in \mathbb{R}$ and $\sigma > 0$. If $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Z \sim \mathcal{N}(0, 1)$, then

$$\frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1), \qquad \text{and} \qquad \sigma Z + \mu \sim \mathcal{N}(\mu, \sigma^2).$$

Proof

We can prove this via a change of variable in the integral for the cumulative distribution function: see Exercise 6.15. A shorter proof goes via the moment generating function, which will be introduced later.

Corollary

If $X \sim \mathcal{N}(\mu, \sigma^2)$ then

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

Try it out

Suppose $X \sim \mathcal{N}(2, 4)$. Find $P(X \geq 5.28)$.
Answer:
We compute

$$\begin{aligned} P(X \geq 5.28) &= P\left(\frac{X-2}{2} \geq \frac{5.28-2}{2}\right) \\ &= P(Z \geq 1.64)\,, \end{aligned}$$

where $Z \sim \mathcal{N}(0, 1)$. Hence

$$P(X \geq 5.28) = 1 - \Phi(1.64) = 1 - 0.95 = 0.05.$$

Textbook references

For more help with this section, check out:

- Section 3.7 in (Blitzstein and Hwang 2019);
- Section 5.2 in (Anderson, Seppäläinen, and Valkó 2018);
- or Section 7.2 in (Stirzaker 2003).

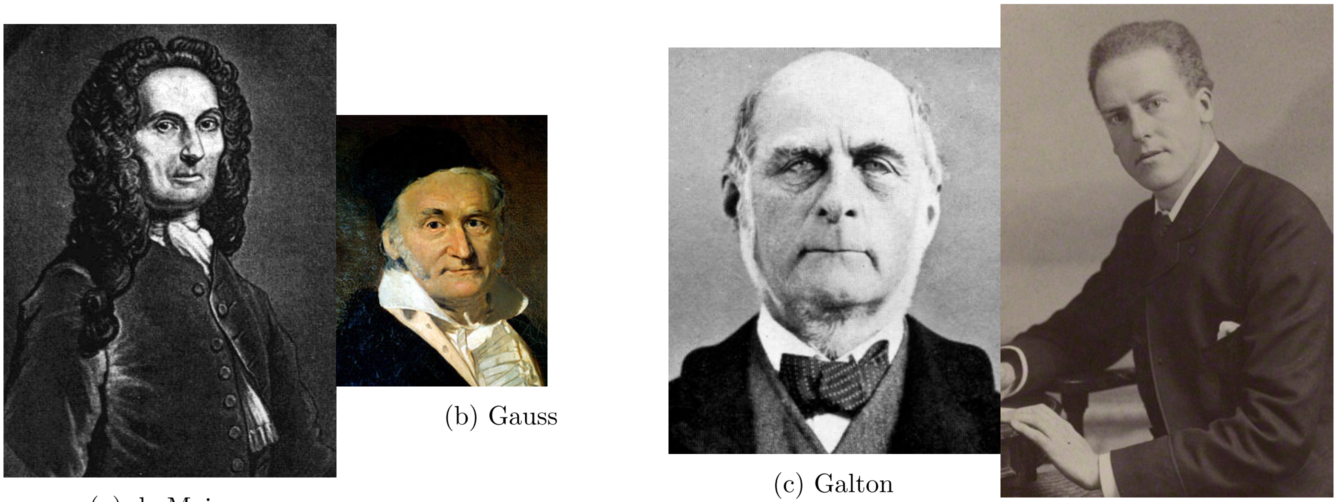
6.12 Historical context

The normal distribution appeared already in work of de Moivre, and is sometimes known as the *Gaussian* distribution after Carl Friedrich Gauss (1777–1855). The name ‘normal distribution’ was applied by eugenicist and biometrician Sir Francis Galton (1822–1911) and statistician Karl Pearson (1857–1936) to mark the distribution’s ubiquity in biometric data.

There is a great deal of subtle and interesting mathematics on the subject of what functions are integrable over what sets. You may see some of this in the third year probability course. The Riemann integral that we use here is sufficient for integrating piecewise continuous functions over finite unions of intervals. Here, we will only consider continuous random variables which have a piecewise continuous probability density function. Other approaches to integration are required to deal with more general functions.

For instance, for infinite countable unions of intervals, we would need the Lebesgue integral (see for instance (Rosenthal 2007)). More precisely, $f(\cdot)$ still determines the value of $P(X \in B)$ when B is an infinite countable union of intervals, but that value is not necessarily given by the Riemann integral.

The treatment of discrete and continuous random variables separately is a little irksome. A general



(a) de Moivre

(b) Gauss

(c) Galton

(d) Pearson

Figure 6.1: (*left to right*) de Moivre, Gauss, Galton, and Pearson.

treatment of random variables, which covers both cases, as well as cases that are neither discrete nor continuous, in a unified setting, requires the mathematical framework of *measure theory*; you will see some of this if you take later probability courses.

7 Multiple random variables

Goals

1. Understand a multivariate random variable as a function from the sample space to a higher dimensional space.

2. Understand jointly distributed discrete random variables:

- their joint probability mass function, marginal probability mass functions, and conditional probability mass functions.
- the partition theorem for discrete random variables.
- independence, and how to apply it.
- the properties of and links between the different probability mass functions, and how those properties and links arise from the axioms of probability distributions.

3. Understand continuously distributed random variables:

- their joint probability density function, marginal probability density functions, and conditional probability density functions.
- the partition theorem for jointly continuously distributed random variables.
- independence, and how to apply it.
- the properties of and links among the different probability density functions.

4. Understand how to work with functions of multiple random variables.

7.1 Joint probability distributions

It is essential for most useful applications of probability to have a theory which can handle many random variables simultaneously. To start, we consider having two random variables. The theory for more than two random variables is an obvious extension of the bivariate case covered below.

Remember that, formally, random variables are simply mappings from Ω into some set. A *bivariate* random variable is a mapping from Ω into a Cartesian product of two sets, i.e., a random variable whose values are ordered pairs of the form (x, y). Of course, a bivariate random variable is a random variable according to our original definition, just with a special kind of set of possible values. However, the concept of bivariate random variable is a useful one if the individual components of the bivariate variable have their own meaning or interest.

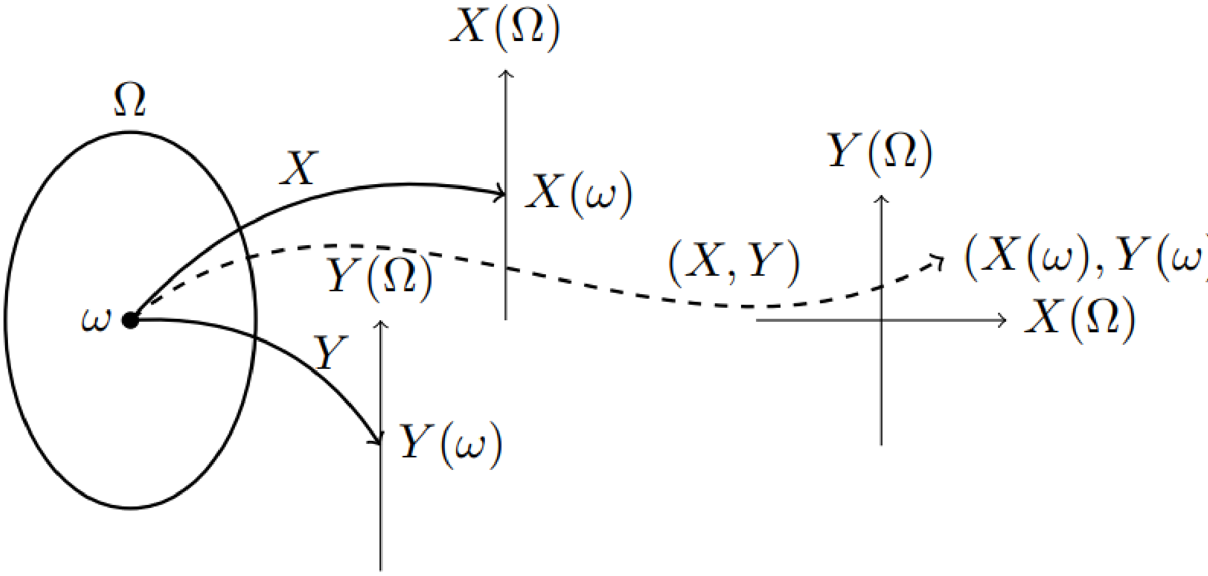
Definition: bivariate random variable

Consider random variables X and Y defined on the same sample space Ω, $X\colon \Omega \to X(\Omega)$ and

$Y\colon \Omega \to Y(\Omega)$. The mapping $(X, Y)\colon \Omega \to (X, Y)(\Omega)$ defined by

$$(X, Y)(\omega) := (X(\omega), Y(\omega))$$

is then a *bivariate random variable*.



Here is a picture:

Note that the set of possible values $(X, Y)(\Omega) = \{(X(\omega), Y(\omega)) : \omega \in \Omega\}$ is a subset of the Cartesian product $X(\Omega) \times Y(\Omega)$.

Example

On sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$, define random variables X and Y by

ω	1	2	3	4	5	6
$X(\omega)$	0	0	0	1	1	1
$Y(\omega)$	0	1	0	2	0	3

Then the bivariate random variable (X, Y) is given by

ω	1	2	3	4	5	6
$(X, Y)(\omega)$	(0,0)	(0,1)	(0,0)	(1,2)	(1,0)	(1,3)

Note that $X(\Omega) = \{0, 1\}$, $Y(\Omega) = \{0, 1, 2, 3\}$, and $(X, Y)(\Omega) = \{(0, 0), (0, 1), (1, 0), (1, 2), (1, 3)\}$ which is a strict subset of $\{0, 1\} \times \{0, 1, 2, 3\}$ (the outcome $(0, 3)$ does not appear, for example).

Similarly to before, for any $A \subseteq X(\Omega) \times Y(\Omega)$, we write '$(X, Y) \in A$' to denote the event

$$\{\omega \in \Omega\colon (X(\omega), Y(\omega)) \in A\}.$$

For any $x \in X(\Omega)$ and $y \in Y(\Omega)$, we write '$X = x, Y = y$' to mean the event $(X, Y) \in \{(x, y)\}$. We sometimes also write $\{(X,Y)=(x,y)\}$ and $\{(X,Y)\in A\}$ to emphasize that these are sets:

$$\begin{aligned}\{X=x,Y=y\} &:= \{(X,Y)\in\{(x,y)\}\}\\ &= \{X=x\}\cap\{Y=y\}\\ &= \{\omega\in\Omega\colon X(\omega)=x \text{ and } Y(\omega)=y\}.\end{aligned}$$

We may also write more complex expressions like:

$$\{0\le X\le Y^2\le 1\}=\{\omega\in\Omega\colon 0\le X(\omega)\le Y(\omega)^2\le 1\}.$$

Key idea: Definition: Independence of two random variables

Two random variables X and Y on the same sample space Ω are *independent* if

$$P(X\in A,\,Y\in B)=P(X\in A)\,P(Y\in B)\ \text{ for all } A\subseteq X(\Omega) \text{ and } B\subseteq Y(\Omega).$$

In other words, X and Y are independent (as random variables) if and only if $\{X\in A\}$ and $\{Y\in B\}$ are independent *as events* for all sets A and B.

Three random variables, X, Y, and Z, say, are independent if the events $\{X\in A\}$, $\{Y\in B\}$, $\{Z\in C\}$ are *mutually* independent. Similarly for any finite collection of random variables. This definition is a bit unwieldy: it reduces to simpler statements in the cases of pairs of discrete or continuous random variables, which we look at next.

Advanced content

Later (when we talk about limit theorems such as the law of large numbers) we will need to talk about infinite sequences of independent random variables. A (possibly infinite) collection of random variables X_i, $i\in\mathcal{I}$, is independent if every *finite* nonempty sub-collection $\mathcal{J}\subseteq\mathcal{I}$ is independent. In other words, X_i, $i\in\mathcal{I}$, are independent if for every finite $\mathcal{J}\subseteq\mathcal{I}$,

$$P\left(\bigcap_{j\in\mathcal{J}}\{X_j\in A_j\}\right)=\prod_{j\in\mathcal{J}}P(X_j\in A_j)\,.$$

7.2 Jointly distributed discrete random variables

Key idea: Key definition: joint probability mass function

Let (X,Y) be a bivariate discrete random variable with $P((X,Y)\in\mathcal{Z})=1$ for a finite or countable $\mathcal{Z}\subseteq(X,Y)(\Omega)$. The *joint probability mass function* $p(\cdot)$ of X and Y is defined by

$$p(x,y):=P(X=x,Y=y)\qquad\text{for all } (x,y)\in\mathcal{Z}.$$

Note there is nothing really new here, other than the terminology: this is just the definition of a discrete random variable written for the special case of a random variable (X,Y) whose values are ordered pairs of the form (x,y).

To avoid ambiguity, we sometimes write if $p_{X,Y}(\cdot)$ if it is not clear from the context which random variables the two arguments refer to; note that $p_{X,Y}(\cdot)$ is not the same as $p_{Y,X}(\cdot)$.

It is the case that (X, Y) is discrete if and only if the individual random variables X and Y are discrete. Proving this is the purpose of the following theorem, which and also explains how the *marginal probability mass functions* $p(x)$ and $p(y)$ of the individual random variables X and Y are connected to the joint probability mass function $p(x, y)$. Note that it does no harm to take $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, extending the definition of $p(x, y)$ with extra 0 values if necessary.

Key idea: Theorem: Joint and single discrete random variables

Let X and Y be two random variables on the same sample space Ω. Then the bivariate random variable (X, Y) is discrete if and only if X and Y are both discrete. Moreover, if X and Y are both discrete with $P(X \in \mathcal{X}) = P(Y \in \mathcal{Y}) = 1$ for finite or countable $\mathcal{X}$ and $\mathcal{Y}$, then their marginal probability mass functions are given in terms of the joint probability mass function by

$$p_X(x) = \sum_{y \in \mathcal{Y}} p(x, y) \text{ for all } x \in \mathcal{X}, \quad p_Y(y) = \sum_{x \in \mathcal{X}} p(x, y) \text{ for all } y \in \mathcal{Y}.$$

Proof

First suppose that X and Y are discrete. Then there exist finite or countable sets of values $\mathcal{X}$ and $\mathcal{Y}$ such that $P(X \in \mathcal{X}) = P(Y \in \mathcal{Y}) = 1$. Hence $P(X \in \mathcal{X}, Y \in \mathcal{Y}) = 1$. The bivariate random variable (X, Y) thus has $P((X, Y) \in \mathcal{X} \times \mathcal{Y}) = 1$. Since $\mathcal{X}$ and $\mathcal{Y}$ are finite or countable, the Cartesian product $\mathcal{X} \times \mathcal{Y}$ is also finite or countable. Hence (X, Y) is discrete.

On the other hand, suppose that (X, Y) is discrete. The possible values in $(X, Y)(\Omega)$ are ordered pairs of the form (x, y), and there is a finite or countable set $\mathcal{Z} \subseteq (X, Y)(\Omega)$ such that $P((X, Y) \in \mathcal{Z}) = 1$. But if we set $\mathcal{X} = \{x : (x, y) \in \mathcal{Z}\}$ we have $P(X \in \mathcal{X}) = P((X, Y) \in \mathcal{Z}) = 1$, and $\mathcal{X}$ is finite or countable (check this!), so X is discrete. Similarly for Y. It remains to notice that, for example, with the same definition of $\mathcal{X}$,

$$\begin{aligned} P(Y = y) &= P(X \in \mathcal{X}, Y = y) \\ &= P(\cup_{x \in \mathcal{X}} \{X = x, Y = y\}) \\ &= \sum_{x \in \mathcal{X}} p(x, y), \end{aligned}$$

where we have used **A4**, the fact that $\mathcal{X}$ is countable, and $P(X \in \mathcal{X}) = 1$.

Try it out

Roll two fair six-sided dice. Let X be the number of 6s rolled, and let Y be the number of 1s and 2s. Find the joint probability mass function $p(x, y)$.

Answer:

The best way to present this is in a table:

$p(x, y)$	$x = 0$	$x = 1$	$x = 2$
$y = 0$	$9/36$	$6/36$	$1/36$
$y = 1$	$12/36$	$4/36$	0
$y = 2$	$4/36$	0	0

For example,

$$\begin{aligned}
p(0,0) &= P(\text{both scores in } \{3,4,5\}) = \frac{9}{36}, \\
p(0,1) &= P(\text{first is } 1,2 \text{ and second } 3,4,5) + P(\text{first is } 3,4,5 \text{ and second } 1,2) \\
&= \frac{2\cdot 3}{36} + \frac{3\cdot 2}{36} = \frac{12}{36},
\end{aligned}$$

and so on. Note that $p(x,y)$ sums to 1!

Again, by **C7**, the joint probability mass function determines the joint probability distribution of X and Y, and the values of the joint probability mass function sum to one:

Theorem: probability mass functions determine distributions for multiple random variables

Let X and Y be discrete random variables with $P((X,Y) \in \mathcal{Z}) = 1$ for a finite or countable $\mathcal{Z}$. Then we have that

$$P((X,Y) \in A) = \sum_{(x,y)\in A} p(x,y) \quad \text{for all } A \subseteq \mathcal{Z}. \tag{7.1}$$

In particular,

$$\sum_{(x,y)\in\mathcal{Z}} p(x,y) = 1.$$

This isn't really a new idea: it is just the result about probability mass functions from Section 6.2 rewritten for the case of a discrete random variable whose possible values are ordered pairs (x,y).

Try it out

Continuing with the "two-dice" example, use $p(x,y)$ to find $P(X \geq 1, Y \leq 1)$. Also, compute the marginal probability mass functions $p_X(x)$ and $p_Y(y)$.

Answer:

We have that

$$\begin{aligned}
P(X \geq 1, Y \leq 1) &= P((X,Y) \in \{(1,0),(1,1),(2,0),(2,1)\}) \\
&= p(1,0) + p(1,1) + p(2,0) + p(2,1) \\
&= \frac{6+4+1+0}{36} = \frac{11}{36}.
\end{aligned}$$

For the marginal distributions, we sum down columns and along rows in the table:

$p(x,y)$	$x=0$	$x=1$	$x=2$	$p_Y(y)$
$y=0$	9/36	6/36	1/36	16/36
$y=1$	12/36	4/36	0	16/36
$y=2$	4/36	0	0	4/36
$p_X(x)$	25/36	10/36	1/36	-

Note that this gives the same result as the binomial distribution, since $X \sim \text{Bin}(2, 1/6)$ and $Y \sim \text{Bin}(2, 1/3)$, so, for example,

$$P(X = 1) = \binom{2}{1} \cdot (1/6)^1 \cdot (5/6)^1 = \frac{10}{36}.$$

Examples

1. In a card game played with a standard 52 card deck, hearts are worth 1, the queen of spades is worth 13 and all other cards worth 0. Let X, Y denote the values of the first and second cards dealt (without replacement, as usual). The possible outcomes of the bivariate random variable (X, Y) are
$$\{(0,0),(1,0),(0,1),(1,1),(0,13),(13,0),(1,13),(13,1)\}$$
(not $(13, 13)$). With a well shuffled deck, the event A that the pair does not include the queen of spades, is
$$\{X \leq 1, Y \leq 1\} = \{(X,Y) \in A\}$$
where
$$A = \{(0,0),(1,0),(0,1),(1,1)\}.$$
So, by Equation 7.1, and a few counting arguments,
$$P((X,Y) \in A) = \frac{38}{52}\frac{37}{51} + \frac{1}{4}\frac{38}{51} + \frac{38}{52}\frac{13}{51} + \frac{1}{4}\frac{12}{51} = \frac{51 \times 50}{52 \times 51} = \frac{50}{52}.$$

2. Discrete random variables X and Y are such that X takes possible values 0, 1, 2 while Y takes values 1, 2, 3, 4, and their joint distribution is given by

$p(x,y)$	$y=1$	$y=2$	$y=3$	$y=4$
$x=0$	0	0	0	1/4
$x=1$	0	1/4	1/4	0
$x=2$	1/4	0	0	0

From this table we can calculate $P(X = x) = 1/4,\ 1/2,\ 1/4$ for $x = 0,\ 1,\ 2$ respectively, i.e., $X \sim \text{Bin}(2, 1/2)$. Similarly $P(Y = y) = 1/4$ for $y = 1,\ 2,\ 3,\ 4$ so Y is uniformly distributed on $\{1, 2, 3, 4\}$.

Often we want to know the distribution of one random variable *conditional* on the value of another random variable.

Definition: Conditional probability mass function

Let X and Y be discrete random variables. For $y \in \mathcal{Y}$, the *conditional probability mass function of X given $Y = y$* is defined by

$$p_{X|Y}(x,y) := P(X = x \mid Y = y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$$

for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ such that $p_Y(y) > 0$ and similarly, the *conditional probability mass function of Y given $X = x$* is

$$p_{Y|X}(y, x) := P(Y = y \mid X = x) = \frac{p_{X,Y}(x, y)}{p_X(x)}$$

for all $y \in \mathcal{Y}$ and $x \in \mathcal{X}$ such that $p_X(x) > 0$.

There's nothing new here, apart from the notation: this is just a particular case of the usual definition of conditional probability of one event given another, e.g.,

$$P(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{p_{X,Y}(x, y)}{p_Y(y)}.$$

Try it out

Continuing from the "two dice" examples above, find $p(x|y)$ and calculate the conditional probability $P(X \geq 1 \mid Y = 0)$.
Answer:
Again, this is best presented in a table:

$p_{X\|Y}(x, y)$	$x = 0$	$x = 1$	$x = 2$
$y = 0$	9/16	6/16	1/16
$y = 1$	12/16	4/16	0
$y = 2$	4/4	0	0

For example,

$$\begin{aligned} p_{X|Y}(0, 2) = P(X = 0 \mid Y = 2) &= \frac{P(X = 0, Y = 2)}{P(Y = 2)} \\ &= \frac{p_{X,Y}(0, 2)}{p_Y(2)} \\ &= \frac{4/36}{4/36} = 1. \end{aligned}$$

Hence

$$P(X \geq 1 \mid Y = 0) = p_{X|Y}(1, 0) + p_{X|Y}(2, 0) = \frac{6}{16} + \frac{1}{16} = \frac{7}{16}.$$

There is also a version of **P4** (partition theorem or, law of total probability) for discrete random variables; again, only the notation is new here.

Theorem: partition theorem for discrete random variables

Let X and Y be discrete random variables. Then,

$$p_X(x) = \sum_{y \in \mathcal{Y}} p_{X|Y}(x, y) p_Y(y).$$

Moreover, if X is real-valued, then for any functions $g : \mathcal{Y} \to \mathbb{R}$ and $h : \mathcal{Y} \to \mathbb{R}$ with $g \leq h$, we can write

$$
\begin{aligned}
P(g(Y) \le X \le h(Y)) &= \sum_{y \in \mathcal{Y}} P(g(y) \le X \le h(y) \mid Y = y)\, P(Y = y) \\
&= \sum_{y \in \mathcal{Y}} \sum_{x \in [g(y), h(y)] \cap \mathcal{X}} p_{X|Y}(x, y) p_Y(y).
\end{aligned}
$$

A version of the above result also holds with X and Y swapped.

Recall from the definition of independence of random variables that X and Y are independent random variables if events $\{X \in A\}$ and $\{Y \in B\}$ are independent for all A, B. If X and Y are discrete, the following result gives a simpler characterization of independence.

Key idea: Lemma: Independence of discrete random variables

Two discrete random variables X and Y on the same sample space Ω are independent if and only if

$$p_{X,Y}(x, y) = p_X(x) p_Y(y) \text{ for all } x \in \mathcal{X} \text{ and } y \in \mathcal{Y}.$$

It follows that for independent discrete random variables X and Y, $p_{X|Y}(x, y) = p_X(x)$ whenever $p_Y(y) > 0$, and $p_{Y|X}(y, x) = p_Y(y)$ whenever $p_X(x) > 0$.

Proof

By the definition of independence, we have that

$$P(X = x,\, Y = y) = P(X = x)\, P(Y = y)\,,$$

as required. On the other hand, suppose that $p_{X,Y}(x, y) = p_X(x) p_Y(y)$. Then by the "probability mass functions determine distributions for multiple random variables" theorem, for any sets A and B,

$$
\begin{aligned}
P(X \in A, Y \in B) = P((X, Y) \in A \times B) &= \sum_{x \in A} \sum_{y \in B} p_{X,Y}(x, y) \\
&= \sum_{x \in A} p_X(x) \sum_{y \in B} p_Y(y) \\
&= P(X \in A)\, P(Y \in B)\,,
\end{aligned}
$$

so X and Y are independent.

Try it out

Continuing the dice example, we saw that $P(X = 2, Y = 2) = 0$ but $P(X = 2)\, P(Y = 2) = \frac{1}{36} \cdot \frac{4}{36} \neq 0$. Hence X and Y are *not* independent. To show dependence it is enough to find a single pair (x, y) where the joint probability mass function does not factorize. Conversely, to show independence it is necessary to consider all pairs (x, y).

Textbook references

If you want more help with this section, check out:

- Section 7.1 in (Blitzstein and Hwang 2019);

- or Chapter 6 in (Anderson, Seppäläinen, and Valkó 2018).

7.3 Jointly continuously distributed random variables

Definition: jointly continuous random variables

Consider two real-valued random variables $X : \Omega \to \mathbb{R}$ and $Y : \Omega \to \mathbb{R}$. We say that X and Y are *jointly continuously distributed* when there is a non-negative piecewise continuous function $f(\cdot) : \mathbb{R}^2 \to \mathbb{R}$, called the *joint probability density function*, such that

$$P(X \in [a,b], Y \in [c,d]) = \int_a^b \left(\int_c^d f(x,y)\, dy \right) dx$$

for all $[a,b] \times [c,d] \subseteq \mathbb{R}^2$.

To avoid ambiguity we sometimes write $f_{X,Y}(x,y)$ for the joint probability density of X and Y. The interpretation of $f(x,y)$ is that

$$P(X \in [x, x+dx], Y \in [y, y+dy]) = f(x,y)\, dx\, dy \tag{7.2}$$

for x, y at which $f(x,y)$ is continuous.

As before, the joint probability density function $f(x,y)$ determines the joint probability $P((X,Y) \in A)$ for most events A. More precisely, remember the definition of type I and II regions in the theory of multiple integration: a type I region is of the form

$$D = \{a_0 \leq x \leq a_1,\ \phi_1(x) \leq y \leq \phi_2(x)\}$$

for some continuous functions ϕ_1 and ϕ_2, and a type II region is of the form

$$D = \{\psi_1(y) \leq x \leq \psi_2(y),\ b_0 \leq y \leq b_1\}$$

for some continuous functions ψ_1 and ψ_2.

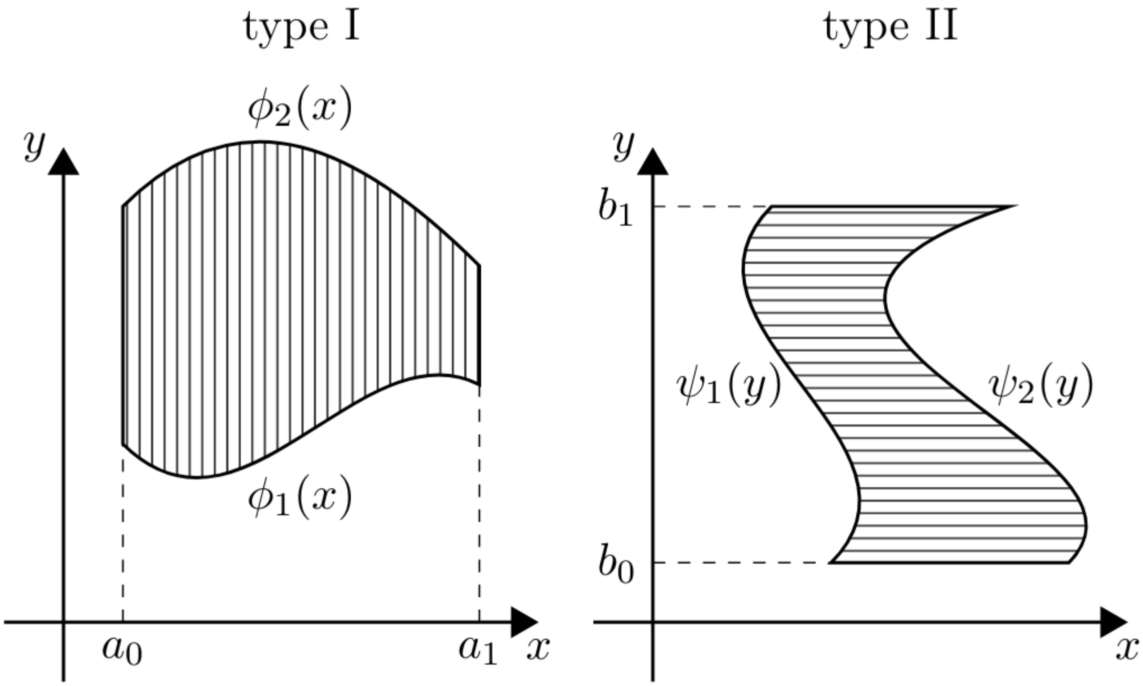


Figure 7.1: Regions of types 1 and 2

As we know from calculus, unions of regions of these types are precisely the regions over which we can integrate[1]. So, we have:

Theorem: probabilities as integrals for joint distributions

If X and Y are jointly continuously distributed, then for any $A \subseteq \mathbb{R}^2$ that is a finite union of type I and type II regions:

$$P((X,Y) \in A) = \iint\limits_A f(x,y)\,dx\,dy.$$

Again as before, $f(x,y)$ integrates to one:

Corollary: joint densities integrate to 1

Let X and Y be jointly continuously distributed random variables. Then their joint probability density function integrates to one:

$$\iint\limits_{\mathbb{R}^2} f(x,y)\,dx\,dy = 1.$$

Try it out

Suppose that X and Y have joint probability density function

$$f(x,y) = c(x^2+y),\ \text{for}\ -1 \le x \le 1,\ 0 \le y \le 1-x^2,$$

[1] At least when restricted to the Riemann integral.

with $f(x, y) = 0$ otherwise. What is the value of c?
Answer: This is an exercise in multiple integration. We have from the Corollary that

$$\begin{aligned}
1 &= \iint\limits_{\mathbb{R}^2} f(x, y)\, dx\, dy \\
&= \int_{-1}^{1} dx \int_{0}^{1-x^2} c(x^2 + y)\, dy \\
&= c \int_{-1}^{1} dx \left[x^2 y + \frac{y^2}{2}\right]_0^{1-x^2} \\
&= \frac{c}{2} \int_{-1}^{1} (1 - x^4)\, dx \\
&= \frac{c}{2} \left[x - \frac{x^5}{5}\right]_{-1}^{1} \\
&= \frac{4}{5} c.
\end{aligned}$$

So we find that $c = 5/4$.

Try it out

Consider random variables X and Y with joint probability density function

$$f(x, y) = \begin{cases} x + y & \text{if } (x, y) \in [0, 1]^2 \\ 0 & \text{otherwise.} \end{cases}$$

(a) Calculate $P(1/4 < X < 3/4, 0 < Y < 1/2)$

(b) Calculate $P(X^2 < Y < X)$

Answer:
Again, this is an exercise in multiple integration. For part (a) we have

$$\begin{aligned}
P(1/4 < X < 3/4, 0 < Y < 1/2) &= \int_{1/4}^{3/4} dx \int_{0}^{1/2} (x + y)\, dy \\
&= \int_{1/4}^{3/4} dx \left[xy + \frac{y^2}{2}\right]_0^{1/2} \\
&= \int_{1/4}^{3/4} \left(\frac{x}{2} + \frac{1}{8}\right) dx \\
&= \frac{3}{16}.
\end{aligned}$$

For part (b), we have

$$P(X^2 < Y < X) = \int_0^1 dx \int_{x^2}^{x} (x+y)\,dy = \int_0^1 dx \left[xy + \frac{y^2}{2}\right]_{x^2}^{x}$$
$$= \int_0^1 \left(\frac{3x^2}{2} - x^3 - \frac{x^4}{4}\right) dx = \frac{3}{20}.$$

Note that, if X and Y are jointly continuously distributed, then for any interval $[a,b]$,

$$P(X \in [a,b]) = P(X \in [a,b], Y \in \mathbb{R}) = \int_a^b \left(\int_{-\infty}^{\infty} f(x,y)\,dy\right) dx,$$

so, by the definition of a continuous random variable, X is continuously distributed as well, as is Y by a similar argument. We have shown the following:

Corollary

Let X and Y be jointly continuously distributed random variables. Then X and Y are (each separately) continuously distributed, with

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy, \text{ for all } x \in \mathbb{R}, \qquad f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx, \text{ for all } y \in \mathbb{R}.$$

In a multivariate context, the probability density functions $f_X(x)$ and $f_Y(y)$ are also called *marginal probability density functions.*

Try it out

Continuing from the previous example, we have that for $x \in [0,1]$,

$$f_X(x) = \int_0^1 (x+y)\,dy = \left[xy + \frac{y^2}{2}\right]_0^1 = x + \frac{1}{2},$$

so

$$f_X(x) = \begin{cases} x + \frac{1}{2} & \text{if } 0 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Try it out

As in a previous example, suppose that X and Y have joint probability density function

$$f(x,y) = c(x^2+y), \text{ for } -1 \le x \le 1,\ 0 \le y \le 1-x^2,$$

with $f(x,y) = 0$ otherwise. we find

$$f_(x) = \frac{5}{4}\int_0^{1-x^2} (x^2+y)\ dy = \frac{5}{8}(1-x^4)$$

for $-1 \leq x \leq 1$ and 0 otherwise;

$$f_(y) = \frac{5}{4}\int_{-\sqrt{1-y}}^{\sqrt{1-y}} (x^2 + y)\,dx = \frac{5}{6}(1 + 2y)\sqrt{1-y}$$

for $0 \leq y \leq 1$ and 0 otherwise.

Advanced content

So, if X and Y are jointly continuously distributed, then both X and Y are also continuously distributed separately. Unlike the discrete case (the "joint and single discrete random variables" theorem), the converse, however, is not true in general, as the following example shows.

Example

Let X be a continuously distributed random variable, and let $Y := 2X$. Then both X and Y are continuously distributed separately, however X and Y are not jointly continuously distributed. Indeed, suppose that X and Y were jointly continuously distributed with density function $f(x, y)$. Then, with $A = \{(x, y) : 2x = y\}$,

$$\iint\limits_A f(x, y)\,dx\,dy = \int_{-\infty}^{+\infty} \left(\int_{2x}^{2x} f(x, y)\,dy \right) dx = 0.$$

This implies that $P(2X = Y) = 0$ for every two jointly continuously distributed random variables X and Y. But, because for our choice of X and Y, obviously $P(2X = Y) = 1$; so, by contradiction, X and Y cannot be jointly continuously distributed.

Conditional probability density function

Let X and Y be jointly continuously distributed random variables. The *conditional probability density function of X at $Y = y$* is defined by

$$f_{X|Y}(x|y) := \frac{f_{X,Y}(x, y)}{f_Y(y)} \qquad \text{for all } x, y \in \mathbb{R} \text{ such that } f_Y(y) > 0.$$

and similarly, the *conditional probability density function of Y at $X = x$* is

$$f_{Y|X}(y|x) := \frac{f_{X,Y}(x, y)}{f_X(x)} \qquad \text{for all } x, y \in \mathbb{R} \text{ such that } f_X(x) > 0.$$

Advanced content

Roughly speaking, $f_{X|Y}(x|y)$ should be thought of as the probability density function of X conditional on $Y = y$. Because the event $Y = y$ has probability 0, this interpretation needs a bit of work to realize rigorously. A formal manipulation using Equation 6.2 and Equation 7.2 goes as follows:

$$\begin{aligned} P(X \in [x, x + dx] \mid Y \in [y, y + dy]) &= \frac{P(X \in [x, x + dx], Y \in [y, y + dy])}{P(Y \in [y, y + dy])} \\ &= \frac{f_{X,Y}(x, y)\, dx\, dy}{f_Y(y)\, dy} = f_{X|Y}(x|y)\, dx, \end{aligned}$$

so $f_{X|Y}(x|y)$ is the probability density of X *conditional on* $Y \in [y, y + dy]$. This relies on $f_{X,Y}$ and f_Y being continuous so that Equation 6.2 and Equation 7.2 are valid.

Theorem: Partition theorem for jointly continuous random variables

Let X and Y be jointly continuously distributed random variables. Then,

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X|Y}(x|y) f_Y(y)\, dy.$$

Moreover, for any piecewise continuous functions $g : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$ with $g \leq h$, we can write,

$$P(g(Y) \leq X \leq h(Y)) = \int_{-\infty}^{+\infty} \left(\int_{g(y)}^{h(y)} f_{X|Y}(x|y)\, dx \right) f_Y(y)\, dy. \tag{7.3}$$

Advanced content

We can formulate this theorem in a way that is more similar to the discrete case. For any event A, and any continuously distributed random variable Y, we define

$$P(A \mid Y = y) := \lim_{h \to 0} P(A \mid Y \in [y, y + h])$$

whenever $P(A \mid Y \in [y, y + h]) > 0$ for all h sufficiently small—this happens when $f_(y)$ is continuous at y and $f_(y) > 0$. Our usual definition of conditional probability does not apply, because $P(Y = y) = 0$, so $P(A \mid Y = y)$ is not really a conditional probability. One should be warned that the notation $P(A \mid Y = y)$ for continuously distributed Y can lead to extremely confusing issues, such as *Borel's paradox*. Anyway, ignoring potential pitfalls, with this notation

$$P(g(y) \leq X \leq h(y) \mid Y = y) = \int_{g(y)}^{h(y)} f(x|y)\, dx,$$

and consequently, we can rewrite Equation 7.3 as

$$P(g(Y) \leq X \leq h(Y)) = \int_{-\infty}^{+\infty} P(g(y) \leq X \leq h(y) \mid Y = y)\, f(y)\, dy.$$

Similarly to the discrete case, independence can be characterized as a factorization property of the joint probability density function.

Key idea: Lemma: independence of jointly continuous random variables

Two jointly continuously distributed random variables X and Y are independent if and only if

$$f_{X,Y}(x,y) = f_X(x) f_Y(y) \text{ for all } x \text{ and } y \in \mathbb{R}.$$

For independent jointly continuously distributed X and Y, $f_{X|Y}(x|y) = f_X(x)$ whenever $f_Y(y) > 0$, and $f_{Y|X}(y|x) = f_Y(y)$ whenever $f_X(x) > 0$.

Example

Suppose X and Y have joint probability density function

$$f(x,y) = \begin{cases} 3e^{-(x+3y)} & \text{if } x \geq 0 \text{ and } y \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Because we can write $f(x,y)$ as $e^{-x} \cdot 3e^{-3y}$ (for $x \geq 0$ and $y \geq 0$), it follows that $X \sim \mathcal{E}(1)$ and $Y \sim \mathcal{E}(3)$, and they are independent. We can calculate things like, for $a > 0$,

$$\begin{aligned} P(aY < X) &= \int_0^\infty \left(\int_{ay}^\infty f(x|y)\, dx \right) f_(y)\, dy \\ &= \int_0^\infty \left(\int_{ay}^\infty e^{-x}\, dx \right) 3e^{-3y}\, dy = \int_0^\infty e^{-ay} \cdot 3e^{-3y}\, dy = 3/(3+a). \end{aligned}$$

Textbook references

If you want more help on this section, check out:

- Section 7.1 in (Blitzstein and Hwang 2019);
- Section 6.2 in (Anderson, Seppäläinen, and Valkó 2018);
- or Sections 8.1 and 8.3 in (Stirzaker 2003).

7.4 Functions of multiple random variables

Suppose $X : \Omega \to X(\Omega)$ and $Y : \Omega \to Y(\Omega)$ are (discrete or continuous) random variables, and $g : X(\Omega) \times Y(\Omega) \to \mathcal{S}$ is some function assigning a value $g(x,y) \in \mathcal{S}$ to each point (x,y). Then $g(X,Y)$ is also a random variable, namely the outcome to a 'new experiment' obtained by running the 'old experiments' to produce values x for X and y for Y, and then evaluating $g(x,y)$.

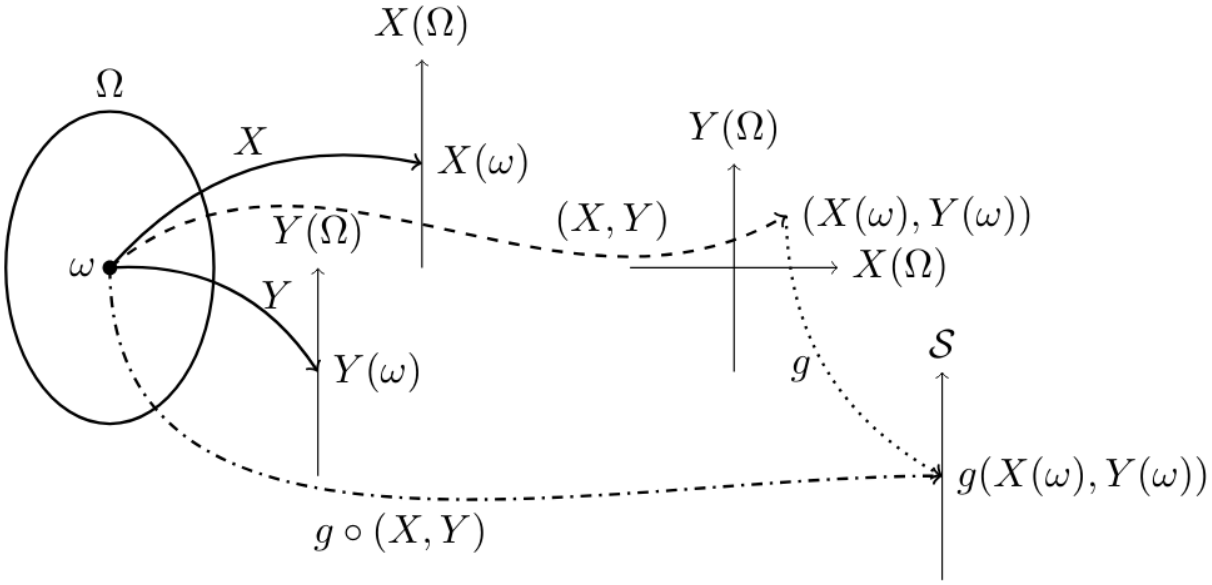


Figure 7.2: $g(X,Y)$ as a random variable

Formally, $g(X,Y) := g\circ(X,Y)$, or in more specific terms, the random variable $g(X,Y):\Omega\to\mathcal{S}$ is defined by:

$$g(X,Y)(\omega) := g(X(\omega),Y(\omega)) \text{ for all } \omega\in\Omega.$$

For example:

$$P(g(X,Y)\in A) = P(\{\omega\in\Omega : g(X(\omega),Y(\omega))\in A\}) \text{ for all } A\subseteq\mathcal{S}.$$

For any random variables X and Y, $X+Y$, $X-Y$, XY, $\min(X,Y)$, $e^{t(X+Y)}$, and so on, are all random variables as well.

Try it out

Consider the jointly continuous random variables X and Y from . Define a new random variable $S = X+Y$.

(a) Calculate $F_S(\cdot)$ and hence deduce that S is a continuous random variable.

(b) Identify f_S.

Answer:
For part (a), we note that $P(0\le S\le 2)=1$ so $F_S(s)=0$ for $s<0$ and $F_S(s)=1$ for $s\ge 2$. Suppose that $0\le s\le 1$. Then

$$\begin{aligned} F_S(s) = P(X+Y\le s) &= \int_0^s\left(\int_0^{s-y}(x+y)\,dx\right)dy \\ &= \int_0^s\left[x^2/2+xy\right]_{x=0}^{x=s-y}\,dy \\ &= \frac{1}{2}\int_0^s(s^2-y^2)\,dy \\ &= \frac{s^3}{3}. \end{aligned}$$

Next, for $1 < s \leq 2$,

$$\begin{aligned}
F_S(s) &= P(X + Y \leq s) \\
&= \int_0^{s-1} \left(\int_0^1 (x + y)\, dx \right) dy + \int_{s-1}^1 \left(\int_0^{s-y} (x + y)\, dx \right) dy \\
&= \int_0^{s-1} (y + 1/2)\, dy + \frac{1}{2} \int_{s-1}^1 (s^2 - y^2)\, dy \\
&= s^2 - \frac{1 + s^3}{3}.
\end{aligned}$$

We see that $F_S(\cdot)$ is continuous, and piecewise differentiable. Hence there is a density which is obtained by differentiation:

$$f_S(s) = \frac{dF_S(s)}{ds} = \begin{cases} s^2 & \text{if } 0 \leq s \leq 1, \\ 2s - s^2 & \text{if } 1 < s \leq 2, \\ 0 & \text{elsewhere,} \end{cases}$$

and in fact $f_S(s)$ is continuous everywhere.

Textbook references

If you want more help on this section, check out:

- Section 3.9 in (DeGroot and Schervish 2013).

8 Expectation

Goals

1. Have an intuitive as well as mathematical understanding of expectation, variance, and covariance. Know how expectation, variance, and covariance, behave under linear transformations and sums.
2. Know how to evaluate expectation of functions of random variables.
3. Know the properties of expectation, variance, and covariance, and the relations between them.
4. Understand the difference between variance and standard deviation.
5. Know conditional expectation, the partition theorem for conditional expectation and the special notation associated with it.
6. Know how expectation, variance, and standard deviation behave under independence, and for sums of independent random variables in particular.
7. Know the Markov and Chebyshev inequalities, where they come from, and how to apply them.

8.1 Definition and interpretation

In a relative frequency interpretation (discussed earlier in Section 4.1), suppose that we run n trials on an experiment where we observe the outcome of some real-valued random variable $X : \Omega \to \mathbb{R}$ in each trial. Let x_i denote the observed value of X in the ith trial; the sequence of observations x_1, x_2, …, x_n is called a *sample*. The *sample mean* is then simply $\frac{1}{n}\sum_{i=1}^{n} x_i$. As a mathematical idealization, we may suppose that there is a unique, empirical limiting value for the sample mean, as n tends to infinity, which we call the *expectation* of X.

In a betting interpretation (discussed earlier in Section 4.2), you can simply consider your 'fair price' for a bet which pays X; that price, we call your *expectation* of X.

The idea of expectation is very interesting mathematically, and also provides ways to use probability in a host of practical applications. Regardless of interpretation, the expectation of X can be connected to the probability mass function $p(x)$ (if X is discrete) or the probability density function $f(x)$ (if X is continuously distributed) in the following way:

Key idea: Definition: expectation

For any real-valued random variable X, the *expectation* (also called *expected value* or *mean*) of X, denoted as $\mathbb{E}[X]$, is defined as:

$$\mathbb{E}[X] := \sum_{x \in \mathcal{X}} x\, p(x) \tag{8.1}$$

if X is discrete, and

$$\mathbb{E}[X] := \int_{-\infty}^{\infty} x \, f(x) \, dx \tag{8.2}$$

if X is continuously distributed, provided that the sum or integral exists.

Examples

1. Suppose that X is discrete with probability mass function

x	1	2
$p(x)$	$\frac{1}{2}$	$\frac{1}{2}$

Then $\mathbb{E}[X] = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 2 = 1.5$.

2. Consider the following ‘game’. You pay Jimmy a pound and then you both throw a fair die. If you get the higher number you get back the difference in pounds, otherwise you lose your pound. Call the return from a game X, with possible values 0, 1, 2, 3, 4 and 5. By counting outcomes,

x	0	1	2	3	4	5
$p(x)$	$\frac{21}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

so that

$$\mathbb{E}[X] = \sum_{x=0}^{5} x \, p(x) = (0 \times 21 + 1 \times 5 + 2 \times 4 + 3 \times 3 + 4 \times 2 + 5 \times 1)/36 = 35/36.$$

Since it costs £1 to play, this means that the expected profit is $-£1/36$. We can interpret this value as meaning that over a long series of games you will get back £35 for every £36 paid out.

3. To find the expectation of a discrete random variable X where $p(x) = 1/n$ for $x \in \{1, 2, \dots, n\}$, we compute

$$\begin{aligned} \mathbb{E}[X] &= \sum_{x=1}^{n} x p(x) \\ &= (1 + 2 + \cdots + n)\frac{1}{n} = \frac{n+1}{2}. \end{aligned}$$

4. If $X \sim \mathrm{U}(a, b)$ then

$$\mathbb{E}[X] = \int_a^b \frac{x}{b-a} \, dx = \left[\frac{x^2/2}{b-a} \right]_a^b = \frac{a+b}{2}.$$

5. If $Z \sim \mathcal{N}(0, 1)$ then

$$\mathbb{E}[Z] = \int_{-\infty}^{+\infty} z\phi(z) \, \mathrm{d}z = 0,$$

since the integrand is an odd function ($\phi(z) = \phi(-z)$).

Try it out

Find the expectation of a continuous random variable X where

$$f(x) = \begin{cases} x/2 & \text{if } x \in [0,2], \\ 0 & \text{elsewhere.} \end{cases}$$

Answer: We compute

$$\mathbb{E}[X] = \int_0^2 x \cdot \frac{x}{2}\,\mathrm{d}x = \left[\frac{x^3}{6}\right]_0^2 = \frac{4}{3}.$$

Advanced content

If the range of possible values for a random variable X is unbounded, then the sum or integral in may fail to exist. In this case, the preceding formulas may still be used to assign a meaningful expectation in some cases, provided we interpret them with care.

For example, if X is discrete with probability mass function $p(x)$, consider

$$\mathbb{E}[X] = \sum_{x \in X(\Omega)} xp(x) = \underbrace{\sum_{x \in X(\Omega): x \geq 0} xp(x)}_{S_+} + \underbrace{\sum_{x \in X(\Omega): x \leq 0} xp(x)}_{S_-};$$

now the individual sums S_+ and S_- *always* exist, but may be *equal to* $+\infty$.

In fact, if we write $X^+ = \max(X,0)$ and $X^- = \max(-X,0)$, then $X = X^+ - X^-$ and $\mathbb{E}[X^+] = S_+$ and $\mathbb{E}[X^-] = S_-$.

To see this, note for example that X^+ is a random variable with $p_{X^+}x = p_X(x)$ for $x > 0$ and $p_{X^+}(0) = P(X \leq 0)$, but only the positive terms contribute to $\mathbb{E}[X^+]$.

It makes sense to say that $\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-]$ (being possibly $-\infty$ or $+\infty$) as long as *at most one* of S_+ and S_- are infinite, using the rules $\infty - x = \infty$ and $x - \infty = -\infty$ for finite x. (There is no sensible interpretation of $\infty - \infty$.) A similar argument applies in the continuous case, with integrals instead of sums. This is summarized in the following table, which shows the values of $\mathbb{E}[X]$ in each case.

	$\mathbb{E}[X^+] < \infty$	$\mathbb{E}[X^+] = \infty$
$\mathbb{E}[X^-] < \infty$	$\mathbb{E}[X^+] - \mathbb{E}[X^-]$	$+\infty$
$\mathbb{E}[X^-] = \infty$	$-\infty$	undefined

Examples

Suppose that X is discrete with probability mass function $p(x) = c_\alpha x^{-\alpha}$ for $x \in \{1, 2, \ldots\}$. This is only a proper probability mass function if $\zeta(\alpha) := \sum_{x=1}^\infty x^{-\alpha} < \infty$, so we need $\alpha > 1$. Then the normalizing constant must be $c_\alpha = 1/\zeta(\alpha)$. But $\mathbb{E}[X] = c_\alpha \sum_{x=1}^\infty x^{1-\alpha}$. If $\alpha \in (1, 2]$, this sum diverges, so $\mathbb{E}[X] = +\infty$. This is the case if, for instance, $p(x) = (6/\pi^2)x^{-2}$.

Textbook references

If you want more help with this section, check out:

- Sections 4.1 and 5.1 in (Blitzstein and Hwang 2019);

- Section 3.3 in (Anderson, Seppäläinen, and Valkó 2018);
- or Sections 4.3 and 7.4 in (Stirzaker 2003).

8.2 Expectation of functions of random variables

Let X be a discrete random variable with $P(X \in \mathcal{X}) = 1$ for a finite or countable set $\mathcal{X}$, and let $g : \mathcal{X} \to \mathbb{R}$ be a real-valued function. As seen in Section 6.11, $g(X) := g \circ X$ is again a random variable. Indeed, $g(X)$ is discrete, since $P(g(X) \in g(\mathcal{X})) = 1$ where $g(\mathcal{X}) := \{g(x) : x \in \mathcal{X}\}$ is finite or countable, and $g(X)$ is a real-valued random variable, so we can define its expectation.

To find the expectation of $g(X)$, by Equation 8.1, according to the definition we need to find the probability mass function $p_{g(X)}()$ first. It turns out however that we can express $\mathbb{E}[g(X)]$ directly in terms of $p_X()$, saving us the effort of having to calculate $p_{g(X)}()$ from $p_X()$.

For any $y \in g(\mathcal{X})$,

$$p_{g(X)}(y) = P(g(X) = y) = \sum_{x \in \mathcal{X}} P(g(X) = y \mid X = x)\, P(X = x) = \sum_{x \in \mathcal{X} : g(x) = y} p(x), \tag{8.3}$$

since

$$P(g(X) = y \mid X = x) = \begin{cases} 1 & \text{if } y = g(x), \\ 0 & \text{otherwise.} \end{cases}$$

It follows that

$$\begin{aligned} \mathbb{E}[g(X)] &= \sum_{y \in g(\mathcal{X})} y\, p_{g(X)}(y) = \sum_{y \in g(\mathcal{X})} y \left(\sum_{x \in \mathcal{X} : g(x) = y} p(x) \right) \\ &= \sum_{y \in g(\mathcal{X})} \left(\sum_{x \in \mathcal{X} : g(x) = y} y p(x) \right) = \sum_{x \in \mathcal{X}} \left(\sum_{y \in g(\mathcal{X}) : y = g(x)} y p(x) \right) \\ &= \sum_{x \in \mathcal{X}} \left(\sum_{y \in g(\mathcal{X}) : y = g(x)} y \right) p(x) = \sum_{x \in \mathcal{X}} g(x) p(x), \end{aligned}$$

where we applied the definition of expectation, Equation 8.3, distributivity, change of order of summation, and distributivity again. A similar result can be proven when X is continuously distributed. Concluding, we have the following result, which is sometimes known as the *Law of the Unconscious Statistician*:

Key idea: Theorem: expectation of a function of a random variable

For any discrete random variable X taking values in $\mathcal{X}$, and any function $g : \mathcal{X} \to \mathbb{R}$,

$$\mathbb{E}[g(X)] = \sum_{x \in \mathcal{X}} g(x) p(x), \tag{8.4}$$

provided that the sum exists. Similarly, for any continuous random variable X and any function $g : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) \,\mathrm{d}x, \tag{8.5}$$

provided that the integral exists.

Examples

1. Suppose that X takes values $0, 1, 2, 3, 4$ each with probability $1/5$. Then

$$\begin{aligned}\mathbb{E}[(X-3)^2] &= \sum_{x=0}^{4}(x-3)^2 p(x)\\ &= \frac{1}{5}((0-3)^2+(1-3)^2+(2-3)^2+(3-3)^2+(4-3)^2)\\ &= \frac{1}{5}(9+4+1+0+1) = 3.\end{aligned}$$

2. Suppose X takes values $-2, -1, 0, 1, 2, 3$ each with probability $1/6$. Then

$$\begin{aligned}\mathbb{E}[X^2] &= \frac{1}{6}((-2)^2+(-1)^2+0+1+2^2+3^2) = \frac{19}{6};\\ \mathbb{E}[\sin(\pi X/4)] &= \frac{1}{6}(-1-1/\sqrt{2}+0+1/\sqrt{2}+1+1/\sqrt{2}) = \frac{1}{6\sqrt{2}};\end{aligned}$$

   and so on.

3. If $X \sim \mathrm{U}(-1, 1)$ then $f(x) = \frac{1}{2}$ for $x \in [-1, 1]$, and zero elsewhere, so

$$\begin{aligned}\mathbb{E}[X^2] &= \int_{-\infty}^{\infty} x^2 f(x)\,\mathrm{d}x\\ &= \int_{-1}^{1} x^2 \cdot \frac{1}{2}\,\mathrm{d}x\\ &= \left[\frac{x^3}{6}\right]_{-1}^{1} = \frac{1}{3}.\end{aligned}$$

4. Note that although $g(X)$ is discrete if X is discrete, if X is continuous then $g(X)$ need not be continuous: for example, if $X \sim \mathrm{U}(0, 2)$ and $g(x) = 1$ if $x \in (0, 1)$ and $g(x) = 0$ otherwise, we have that $g(X)$ is discrete with $P(g(X) = 1) = 1/2$ and $P(g(X) = 0) = 1/2$. In this case $f(x) = 1/2$ for $x \in (0, 2)$, and says that

$$\mathbb{E}[g(X)] = \frac{1}{2}\int_0^2 g(x)\,\mathrm{d}x = \frac{1}{2},$$

   as we would get from a direct calculation for the discrete random variable $g(X)$ as $\mathbb{E}[g(X)] = \frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 1 = \frac{1}{2}$.

Advanced content

Similar comments apply here about extensions of $\mathbb{E}[g(X)]$ to include $+\infty$ or $-\infty$ as at the end of the previous section.

Example

Suppose $X \sim \mathrm{U}(-1,1)$ i.e., X is uniformly distributed on the interval $(-1,1)$, and we set $g(x) = 1/x$ for $x \neq 0$ and $g(0) = 0$. Then $\mathbb{E}[g(X)]$ is not defined because $\mathbb{E}[g(X)^+] = \int_{-1}^{0} 0 \, \frac{1}{2} \, dx + \int_0^1 \frac{1}{2x} \, dx = \infty$ and similarly $\mathbb{E}[g(X)^-] = \infty$.

For multiple random variables, the Law of the Unconscious Statistician reads as follows:

Key idea: Theorem: Expectation of a function of a multivariate random variable

For any discrete random variables X and Y taking values in $\mathcal{X}$ and $\mathcal{Y}$, and any function $g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$,

$$\mathbb{E}[g(X,Y)] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} g(x,y) p(x,y),$$

provided that the sum exists. Similarly, for any jointly continuously distributed random variables X and Y, and any function $g : \mathbb{R}^2 \to \mathbb{R}$,

$$\mathbb{E}[g(X,Y)] = \iint\limits_{\mathbb{R}^2} g(x,y) f(x,y) \, \mathrm{d}x \, \mathrm{d}y,$$

provided that the integral exists.

Examples

1. Consider discrete random variables X and Y with joint probability mass function:

$p(x,y)$	$x=1$	$x=2$	$x=3$
$y=1$	$1/2$	0	$1/8$
$y=2$	0	$1/4$	$1/8$

Then
$$\begin{aligned}\mathbb{E}[(X-2)Y] &= \sum_x \sum_y (x-2)yp(x,y) \\ &= (1-2)\cdot 1\cdot \frac{1}{2} + (2-2)\cdot 2\cdot \frac{1}{4} + (3-2)\cdot 1\cdot \frac{1}{8} + (3-2)\cdot 2\cdot \frac{1}{8} \\ &= -\frac{1}{8}.\end{aligned}$$

2. Consider discrete random variables X and Y with joint probability mass function:

$p(x,y)$	$x=-1$	$x=0$	$x=1$
$y=0$	$1/4$	0	$1/4$
$y=1$	0	$1/4$	$1/4$

Then $\mathbb{E}[XY] = \frac{1}{4}((-1)\times 0 + 0\times 1 + 1\times 0 + 1\times 1) = 1/4$.

Try it out

Let X and Y be jointly continuously distributed random variables, with
$$f(x,y) = \begin{cases} 1 & \text{if } (x,y)\in[0,1]^2, \\ 0 & \text{otherwise.}\end{cases}$$

Find $\mathbb{E}[XY]$.
Answer:
Using the theorem, $\mathbb{E}[XY] = \int_0^1\int_0^1 xy\,\mathrm{d}x\,\mathrm{d}y = (\int_0^1 x\,\mathrm{d}x)(\int_0^1 y\,\mathrm{d}y) = (1/2)^2 = 1/4$.

Textbook references

If you want more help with this section, check out:

- Sections 4.5 and 5.1 in (Blitzstein and Hwang 2019);
- Section 3.3 in (Anderson, Seppäläinen, and Valkó 2018);
- or Sections 4.5, 5.3, 7.4, and 8.5 in (Stirzaker 2003).

8.3 Linearity of expectation

Remember that summation and integration are linear operators, i.e.,
$$\sum_i \alpha f(x_i) + \beta g(x_i) = \alpha \sum_i f(x_i) + \beta \sum_i g(x_i),$$
and
$$\int_A (\alpha f(x) + \beta g(x))\,\mathrm{d}x = \alpha \int_A f(x)\,\mathrm{d}x + \beta \int_A g(x)\,\mathrm{d}x.$$

Consequently,

Key idea: Theorem: linearity of expectation 1

For any real-valued random variable X, and any constants α and $\beta \in \mathbb{R}$,

$$\mathbb{E}[\alpha X + \beta] = \alpha \mathbb{E}[X] + \beta.$$

A similar, but deeper, result is the following.

Key idea: Theorem: linearity of expectation 2

For any two real-valued random variables X and Y on the same sample space Ω,

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

More generally, for any real-valued random variables X_1, X_2, ..., X_n,

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i].$$

Proof

We give the proof in the case where X and Y are discrete. Consider the multiple random variable (X, Y) and the function $g(x, y) = x + y$. By the Law of the Unconscious Statistician we get

$$\begin{aligned}\mathbb{E}[g(X,Y)] &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}(x+y)p(x,y) = \sum_{x\in\mathcal{X}} x \sum_{y\in\mathcal{Y}} p(x,y) + \sum_{y\in\mathcal{Y}} y \sum_{x\in\mathcal{X}} p(x,y)\\ &= \sum_{x\in\mathcal{X}} x p_X(x) + \sum_{y\in\mathcal{Y}} y p_Y(y) = \mathbb{E}[X] + \mathbb{E}[Y],\end{aligned}$$

as claimed. A similar calculation applies in the jointly continuous case, and the extension to more than two random variables follows by induction.

Try it out

Suppose that $X \sim \text{Bin}(n, p)$. What is $\mathbb{E}[X]$?
Answer: We could use the probability mass function and compute

$$\mathbb{E}[X] = \sum_{x=0}^{n} x p(x) = \sum_{x=0}^{n} \binom{n}{x} x p^x (1-p)^{n-x},$$

but now some work is needed to evaluate this (exercise!).
Here is a neater way that will also be useful later on. Recall that X counts the number of success on n independent trails. If we let $Y_i = 1$ if trial i is a success and $Y_i = 0$ if trial i is a failure, then in the binomial scenario $Y_1, \dots, Y_n$ are independent with $P(Y_i = 1) = p$ and $P(Y_i = 0) = 1 - p$. In other words, we may write

$$X = Y_1 + Y_2 + \dots + Y_n,$$

where $Y_i \sim \text{Bin}(1, p)$ are independent Bernoulli random variables. Then $\mathbb{E}[Y_i] = p$ and so $\mathbb{E}[X] = \mathbb{E}[Y_1 + \dots + Y_n] = np$. Note that the independence of the trials is not necessary for this result.

Advanced content

We have proved linearity of expectation for discrete and jointly continuous random variables, but linearity of expectation is true for *all* random variables, and follows from the general measure-theoretic approach to probability theory, in which expectation is defined as a *Lebesgue integral* with respect to a probability measure; this includes the discrete and continuous settings we study as special cases. Students who take later probability courses will see some of this general approach. Note that without measure theory it is hard to prove the theorem in the case where say X is discrete but Y is continuous. If X and Y are *independent*, and at least one of them is continuous, then the sum $X + Y$ is also continuous: this is a theorem (see, e.g. (Moran 1968 Theorem 5.9, p.230)) and an example can be seen in Exercise 6.31, but this may fail without independence.

Textbook references

If you want more help with this section, check out:

- Section 4.2 in (Blitzstein and Hwang 2019);
- Sections 4.2 and 4.6 in (DeGroot and Schervish 2013);
- or Section 5.3 in (Stirzaker 2003).

8.4 Variance and covariance

As mentioned earlier, we can interpret the expectation of X as a long-run average of a sample from distribution X. A popular and mathematically convenient way to measure the variability of X—i.e. to measure how much X varies from $\mathbb{E}[X]$ in the long run—goes via the expectation of the random variable $(X - \mathbb{E}[X])^2$.

Key idea: Definition: Variance

Let X be any real-valued random variable. The *variance* of X is defined as

$$\mathrm{Var}\,(X) := \mathbb{E}\left[(X - \mathbb{E}[X])^2\right],$$

and the *standard deviation* of X is defined as

$$\sigma\,(X) := \sqrt{\mathrm{Var}\,(X)}.$$

Note that both $\mathrm{Var}\,(X)$ and $\sigma\,(X)$ are non-negative numbers.

Using LOTUS, we can immediately derive the following expressions for the variance:

$$\begin{aligned} \mathrm{Var}\,(X) &= \sum_{x \in \mathcal{X}} (x - \mathbb{E}[X])^2 p(x) && \text{if } X \text{ is discrete, and} \\ \mathrm{Var}\,(X) &= \int_{-\infty}^{\infty} (x - \mathbb{E}[X])^2 f(x)\,\mathrm{d}x && \text{if } X \text{ is continuously distributed,} \end{aligned}$$

provided that the sum or integral exists.

Try it out

As in a previous example, suppose that X takes values $0, 1, 2, 3, 4$ each with probability $1/5$. What is $\mathrm{Var}\,(X)$?

Answer:

First we need to compute $\mathbb{E}[X]$, so

$$\mathbb{E}[X] = \sum_{x=0}^{4} xp(x) = \frac{0+1+2+3+4}{5} = \frac{10}{5} = 2.$$

Then

$$\begin{aligned}\mathrm{Var}\,(X) &= \sum_{x=0}^{4} (x - \mathbb{E}[X])^2 p(x) \\ &= \frac{(-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2}{5} = 2.\end{aligned}$$

Examples

1. If X takes values 0, 10, 20 each with probability $1/3$, then

   $$\mathbb{E}[X] = 0 \times \frac{1}{3} + 10 \times \frac{1}{3} + 20 \times \frac{1}{3} = 10.$$

   Consequently, using this value,

   $$\mathrm{Var}\,(X) = \frac{1}{3} \times \left((0-10)^2 + (10-10)^2 + (20-10)^2\right) = \frac{200}{3},$$

   and so $\sigma\,(X) = \sqrt{\frac{200}{3}} \approx 8.16$.

2. Let $Z \sim \mathcal{N}(0,1)$. We know from an earlier example that $\mathbb{E}[Z] = 0$ and by Exercise 8.10, $\mathbb{E}[Z^2] = 1$. Consequently,

   $$\mathrm{Var}\,(Z) = \mathbb{E}[(Z - E(Z)^2)] = \mathbb{E}[Z^2] = 1,$$

   and $\sigma\,(Z) = \sqrt{\mathrm{Var}\,(Z)} = 1$.

For two real-valued random variables, we can ask ourselves how they vary jointly.

Key idea: Definition: covariance

Let X and Y be two real-valued random variables on the same sample space. The *covariance* of X and Y is defined as

$$\mathrm{Cov}\,(X,Y) := \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

We also use the following qualitative terminology.

- If $\mathrm{Cov}\,(X,Y) > 0$ it means that $X - \mathbb{E}[X]$ and $Y - \mathbb{E}[Y]$ tend to have the same sign. That is, if $X > \mathbb{E}[X]$ then it tends to be the case that $Y > \mathbb{E}[Y]$ (or, conversely, if $X < \mathbb{E}[X]$ then it tends to be the case that $Y < \mathbb{E}[Y]$ too). In this case we say that X and Y are *positively correlated.*
- If $\mathrm{Cov}\,(X,Y) < 0$ we say that X and Y are *negatively correlated.* Now $X - \mathbb{E}[X]$ and $Y - \mathbb{E}[Y]$ tend

to have opposite signs.

- If $\mathrm{Cov}\left(X,Y\right)=0$ we say that X and Y are *uncorrelated*.

Note: uncorrelated is *not* the same as independent (more on this later). A quantification of the correlation is provided by the *correlation coefficient*, given by

$$\rho(X,Y):=\frac{\mathrm{Cov}\left(X,Y\right)}{\sqrt{\mathrm{Var}\left(X\right)\mathrm{Var}\left(Y\right)}}.$$

It can be proved (see Exercises 8.x and 8.x) that

$$-1\leq\rho(X,Y)\leq 1.$$

We will see an example below where the correlation coefficient is 1.

Using LOTUS for multiple random variables, we immediately derive the following expressions for the covariance:

$$\mathrm{Cov}\left(X,Y\right)=\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}(x-\mathbb{E}[X])(y-\mathbb{E}[Y])p(x,y)$$

if X and Y are discrete, and

$$\mathrm{Cov}\left(X,Y\right)=\iint_{\mathbb{R}^2}(x-\mathbb{E}[X])(y-\mathbb{E}[Y])f(x,y)\,\mathrm{d}x\,\mathrm{d}y,$$

if X and Y are jointly continuously distributed, provided that the double sum or double integral exists.

Try it out

Consider discrete random variables X and Y with distribution given by

$p(x,y)$	$x=1$	$x=2$	$p_Y(y)$
$y=1$	1/4	0	1/4
$y=4$	0	3/4	3/4
$p_X(x)$	1/4	3/4	

Find their expectations and their covariance.
Answer:
Then $\mathbb{E}[X]=1\cdot\frac{1}{4}+2\cdot\frac{3}{4}=\frac{7}{4}$ and $\mathbb{E}[Y]=1\cdot\frac{1}{4}+4\cdot\frac{3}{4}=\frac{13}{4}$. We compute, using the formula for expectation of a function (LOTUS again),

$$\begin{aligned}\mathrm{Cov}\left(X,Y\right)&=\mathbb{E}[\left(X-\frac{7}{4}\right)\left(Y-\frac{13}{4}\right)]\\&=\sum_x\sum_y\left(x-\frac{7}{4}\right)\left(y-\frac{13}{4}\right)p(x,y)\\&=\frac{1}{4}\left(1-\frac{7}{4}\right)\left(1-\frac{13}{4}\right)+\frac{3}{4}\left(2-\frac{7}{4}\right)\left(4-\frac{13}{4}\right)\\&=\frac{9}{16}.\end{aligned}$$

This means that X and Y are positively correlated, which makes sense from the shape of the table.

We note two simple but important properties.

Proposition: Variance as covariance

For any real-valued random variable X,

$$\mathrm{Var}\left(X\right) = \mathrm{Cov}\left(X, X\right).$$

Proposition: Symmetry of covariance

For any real-valued random variables X and Y,

$$\mathrm{Cov}\left(X, Y\right) = \mathrm{Cov}\left(Y, X\right).$$

As immediate consequences of linearity of expectation (see Section 8.3), we obtain the formulæ:

Corollary: Variance and covariance of linear combinations

For any real-valued random variable X, and any constants α and $\beta \in \mathbb{R}$,

$$\mathrm{Var}\left(\alpha + \beta X\right) = \beta^2 \mathrm{Var}\left(X\right).$$

For any real-valued random variables X, Y, and Z, and any constants α, β, γ, and $\delta \in \mathbb{R}$,

$$\mathrm{Cov}\left(\alpha + \beta X, \gamma + \delta Y\right) = \beta\delta \mathrm{Cov}\left(X, Y\right); \tag{8.6}$$

$$\mathrm{Cov}\left(X + Y, Z\right) = \mathrm{Cov}\left(X, Z\right) + \mathrm{Cov}\left(Y, Z\right); \tag{8.7}$$

and

$$\mathrm{Cov}\left(X, Y + Z\right) = \mathrm{Cov}\left(X, Y\right) + \mathrm{Cov}\left(X, Z\right). \tag{8.8}$$

Note: Equation 8.6 through Equation 8.8 mean that $\mathrm{Cov}\left(\right)$ is a *bilinear operator*.

Proof

We give an example of the type of calculation:

$$\begin{aligned}
\mathrm{Var}\left(\alpha + \beta X\right) &= \mathbb{E}[(\alpha + \beta X - \mathbb{E}[\alpha + \beta X])^2] \\
&= \mathbb{E}[(\alpha + \beta X - \alpha - \beta\mathbb{E}[X])^2] \\
&= \mathbb{E}[(\beta X - \beta\mathbb{E}[X])^2] \\
&= \beta^2\mathbb{E}[(X - \mathbb{E}[X])^2] \\
&= \beta^2 \mathrm{Var}\left(X\right).
\end{aligned}$$

The other statements are similar.

Try it out

Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then, $Z := \frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$, and as we saw earlier, $\mathbb{E}[Z] = 0$ and $\mathrm{Var}\left(Z\right) = 1$. Consequently,

$$\begin{aligned}
\mathbb{E}[X] = \mathbb{E}\left[\mu + \sigma Z\right] = \mu + \sigma\mathbb{E}[Z] = \mu, \\
\mathrm{Var}\left(X\right) = \sigma^2 \mathrm{Var}\left(Z\right) = \sigma^2.
\end{aligned}$$

In other words, the parameters μ and σ^2 of a normal distribution correspond to the expectation and variance, respectively.

We also obtain a slightly different way of calculating the variance:

Corollary: Variance and expectation

For any real-valued random variable X,

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

Proof

Observe that, by linearity of expectation,

$$\begin{aligned}\operatorname{Var}(X) &= \mathbb{E}\left[\left(X - \mathbb{E}[X]\right)^2\right] \\ &= \mathbb{E}\left[X^2 - 2X\mathbb{E}[X] + (\mathbb{E}[X])^2\right] \\ &= \mathbb{E}\left[X^2\right] - 2\mathbb{E}[X]\mathbb{E}[X] + (\mathbb{E}[X])^2 \\ &= \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2,\end{aligned}$$

as required.

Example

If X takes values 0, 10, 20 each with probability $1/3$, then

$$\begin{aligned}\mathbb{E}[X] &= 0 \times \frac{1}{3} + 10 \times \frac{1}{3} + 20 \times \frac{1}{3} = 10, \\ \mathbb{E}\left[X^2\right] &= 0^2 \times \frac{1}{3} + 10^2 \times \frac{1}{3} + 20^2 \times \frac{1}{3} = 500/3.\end{aligned}$$

Consequently,

$$\operatorname{Var}(X) = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 = 100 - 500/3 = 200/3,$$

which agrees with the value that we found earlier.

The next example shows how our various formulae can be put to good use.

Try it out

Suppose that X has probability mass function

x	0	3	10
$p(x)$	1/4	1/2	1/4

Define $Y = 2X - 6$. Find

(a) $\mathbb{E}[X]$, $\mathbb{E}\left[X^2\right]$, and $\mathrm{Var}\,(X)$;

(b) $\mathbb{E}[Y]$ and $\mathrm{Var}\,(Y)$;

(c) $\mathrm{Cov}\,(X, Y)$.

Answer:
For (a) we calculate that $\mathbb{E}[X] = 3 \cdot \frac{1}{2} + 10 \cdot \frac{1}{4} = 4$ and $\mathbb{E}\left[X^2\right] = 3^2 \cdot \frac{1}{2} + 10^2 \cdot \frac{1}{4} = \frac{59}{2}$. So $\mathrm{Var}\,(X) = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 = \frac{59}{2} - 4^2 = \frac{27}{2}$.
For (b), we compute swiftly that

$$\mathbb{E}[Y] = \mathbb{E}\left[2X - 6\right] = 2\mathbb{E}[X] - 6 = 2,$$

and

$$\mathrm{Var}\,(Y) = \mathrm{Var}\,(2X - 6) = 4\mathrm{Var}\,(X) = 54.$$

Finally, for (c),

$$\mathrm{Cov}\,(X, Y) = \mathrm{Cov}\,(X, 2X - 6) = 2\mathrm{Cov}\,(X, X) = 2\mathrm{Var}\,(X) = 27.$$

Note that

$$\rho(X, Y) = \frac{\mathrm{Cov}\,(X, Y)}{\sqrt{\mathrm{Var}\,(X)\,\mathrm{Var}\,(Y)}} = 1,$$

which makes sense since X and Y are perfectly positively correlated.

We also obtain a slightly different way of calculating the covariance:

Corollary: Covariance via expectations

For any real-valued random variables X and Y,

$$\mathrm{Cov}\,(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

Proof

The calculation should now be familiar:

$$\begin{aligned}
\mathrm{Cov}\,(X, Y) &= \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right] \\
&= \mathbb{E}\left[XY - Y\mathbb{E}[X] - X\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y]\right] \\
&= \mathbb{E}\left[XY\right] - \mathbb{E}[Y]\mathbb{E}[X] - \mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y] \\
&= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y],
\end{aligned}$$

as required.

Examples

We return to the previous example, with joint pmf given by

$p(x,y)$	$x=1$	$x=2$	$p_Y(y)$
$y=1$	$1/4$	0	$1/4$
$y=4$	0	$3/4$	$3/4$
$p_X(x)$	$1/4$	$3/4$	

We already saw that $\mathbb{E}[X]=7/4$ and $\mathbb{E}[Y]=13/4$. Now we can compute $\mathbb{E}[XY]=\frac{1}{4}\cdot 1+\frac{3}{4}\cdot 8=\frac{25}{4}$, so that $\mathrm{Cov}\left(X,Y\right)=\frac{25}{4}-\frac{7}{4}\cdot\frac{13}{4}=\frac{9}{16}$, as we obtained before.

Try it out

Suppose X takes values 0, 1, 2 with probabilities 1/4, 1/2, 1/4. Let $Y:=X^2$. What is $\mathrm{Var}\left(X\right)$, $\mathrm{Var}\left(Y\right)$, and $\mathrm{Cov}\left(X,Y\right)$?

Answer:

First, $\mathbb{E}[X]=1$.

Next, $Y=X^2$ takes values 0, 1, 4 with probabilities 1/4, 1/2, 1/4, so $\mathbb{E}\left[X^2\right]=\mathbb{E}[Y]=3/2$. Similarly, $Y^2=X^4$ takes values 0, 1, 16 with probabilities 1/4, 1/2, 1/4, so $\mathbb{E}\left[Y^2\right]=9/2$. Finally, $XY=X^3$ takes values 0, 1, 8 with probabilities 1/4, 1/2, 1/4, so $\mathbb{E}[XY]=5/2$. Concluding,

$$\begin{aligned}\mathrm{Var}\left(X\right)&=\mathbb{E}\left[X^2\right]-(\mathbb{E}[X])^2=3/2-1=1/2;\\ \mathrm{Var}\left(Y\right)&=\mathbb{E}\left[Y^2\right]-(\mathbb{E}[Y])^2=9/2-9/4=9/4;\\ \mathrm{Cov}\left(X,Y\right)&=\mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y]=5/2-3/2=1.\end{aligned}$$

Finally, we can now also say something about the variance of sums of random variables:

Theorem: Variance of a sum

For any real-valued random variables X and Y on the same sample space,

$$\mathrm{Var}\left(X+Y\right)=\mathrm{Var}\left(X\right)+\mathrm{Var}\left(Y\right)+2\mathrm{Cov}\left(X,Y\right).$$

More generally, for any real-valued random variables X_1, X_2, ..., X_n,

$$\mathrm{Var}\left(\sum_{i=1}^{n}X_i\right)=\sum_{i=1}^{n}\mathrm{Var}\left(X_i\right)+2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\mathrm{Cov}\left(X_i,X_j\right).$$

Proof

For the first statement,

$$\begin{aligned}\mathrm{Var}\left(X+Y\right)&=\mathrm{Cov}\left(X+Y,X+Y\right)\\ &=\mathrm{Cov}\left(X,X+Y\right)+\mathrm{Cov}\left(Y,X+Y\right)\\ &=\mathrm{Cov}\left(X,X\right)+\mathrm{Cov}\left(X,Y\right)+\mathrm{Cov}\left(Y,X\right)+\mathrm{Cov}\left(Y,Y\right),\end{aligned}$$

which gives the result. In general,

$$\begin{aligned}
\text{Var}\left(\sum_{i=1}^{n} X_i\right) &= \text{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right)\\
&= \sum_{i=1}^{n} \text{Cov}\left(X_i, \sum_{j=1}^{n} X_j\right)\\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \text{Cov}\left(X_i, X_j\right)\\
&= \sum_{i=1}^{n} \text{Cov}\left(X_i, X_i\right) + \sum_{i=1}^{n}\sum_{j\neq i} \text{Cov}\left(X_i, X_j\right)\\
&= \sum_{i=1}^{n} \text{Var}\left(X_i\right) + \sum_{i=1}^{n}\sum_{j=1}^{i-1} \text{Cov}\left(X_i, X_j\right) + \sum_{i=1}^{n}\sum_{j=i+1}^{n} \text{Cov}\left(X_i, X_j\right),
\end{aligned}$$

with the convention that an empty sum is zero.

So, in general, the variance of a sum is *not* equal to the sum of the variances, unless all covariances are zero (zero covariance occurs under independence, covered later).

For example, for three real-valued random variables X, Y, and Z,

$$\text{Var}\left(X+Y+Z\right) = \text{Var}\left(X\right) + \text{Var}\left(Y\right) + \text{Var}\left(Z\right) + 2\left(\text{Cov}\left(X,Y\right) + \text{Cov}\left(X,Z\right) + \text{Cov}\left(Y,Z\right)\right).$$

Try it out

Suppose X takes values 0, 1, 2 with probabilities 1/4, 1/2, 1/4. Let $Y := X^2$.
We already found earlier that $\text{Var}\left(X\right) = 1/2$, $\text{Var}\left(Y\right) = 9/4$, and $\text{Cov}\left(X,Y\right) = 1$. Now also find $\text{Var}\left(X+Y\right)$ and $\text{Var}\left(X-Y\right)$.
Answer:
By the above, $\text{Var}\left(X+Y\right) = \text{Var}\left(X\right) + \text{Var}\left(Y\right) + 2\text{Cov}\left(X,Y\right) = 1/2 + 9/4 + 2\times 1 = 19/4$.
(Note that in this very simple example we can easily calculate $\text{Var}\left(X+Y\right)$ directly, as $X+Y$ takes possible values 0, 2, 6 with probabilities 1/4, 1/2, 1/4 so we can confirm by direct calculation that $\text{Var}\left(X+Y\right) = 19/4$.)
Next, note that $\text{Var}\left(X-Y\right) = \text{Var}\left(X+Z\right)$, where $Z = -Y$. As $\text{Var}\left(Z\right) = (-1)^2\text{Var}\left(Y\right) = \text{Var}\left(Y\right)$ and $\text{Cov}\left(X,Z\right) = \text{Cov}\left(X,-Y\right) = -\text{Cov}\left(X,Y\right)$ we have

$$\begin{aligned}
\text{Var}\left(X-Y\right) &= \text{Var}\left(X\right) + \text{Var}\left(Y\right) - 2\text{Cov}\left(X,Y\right)\\
&= 1/2 + 9/4 - 2\times 1 = 3/4.
\end{aligned}$$

Try it out

Suppose that n people throw their hats high in the air, and each catches a hat which is equally likely to be any of the n hats. Let H be the number of people that catch their own hat. Find $\mathbb{E}\left[H\right]$ and $\text{Var}\left(H\right)$.
Answer:

The distribution of H is quite hard to find, but $\mathbb{E}[H]$ and $\operatorname{Var}(H)$ are fairly straightforward. Define

$$X_i := \begin{cases} 1 & \text{if person } i \text{ catches their own hat,} \\ 0 & \text{otherwise.} \end{cases}$$

Then $H = \sum_{i=1}^n X_i$. Here

$$\mathbb{E}[X_i] = \frac{1}{n} \times 1 + \frac{n-1}{n} \times 0 = \frac{1}{n},$$

and so

$$\mathbb{E}[H] = \mathbb{E}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \mathbb{E}[X_i] = 1.$$

Similarly, because $X_i^2 = X_i$,

$$\operatorname{Var}(X_i) = \mathbb{E}[X_i^2] - (\mathbb{E}[X_i])^2 = \frac{1}{n} - \frac{1}{n^2} = \frac{n-1}{n^2}.$$

We also need $\operatorname{Cov}(X_i, X_j)$. We compute

$$\begin{aligned} \mathbb{E}[X_i X_j] &= P(X_i = 1, X_j = 1) \\ &= P(X_i = 1)\, P(X_j = 1 \mid X_i = 1) \\ &= \frac{1}{n} \cdot \frac{1}{n-1}, \end{aligned}$$

so

$$\begin{aligned} \operatorname{Cov}(X_i, X_j) &= \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\,\mathbb{E}[X_j] \\ &= \frac{1}{n(n-1)} - \frac{1}{n^2} = \frac{1}{n^2(n-1)}. \end{aligned}$$

Hence

$$\begin{aligned} \operatorname{Var}(H) &= \sum_{i=1}^n \operatorname{Var}(X_i) + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \operatorname{Cov}(X_i, X_j) \\ &= n \operatorname{Var}(X_1) + n(n-1) \operatorname{Cov}(X_1, X_2) \\ &= n \times \frac{n-1}{n^2} + n(n-1) \times \frac{1}{n^2(n-1)} \\ &= \frac{n-1+1}{n} = 1. \end{aligned}$$

Textbook references

If you want more help with this section, check out:

- Sections 4.6 and 7.3 in (Blitzstein and Hwang 2019);
- Section 3.4 in (Anderson, Seppäläinen, and Valkó 2018);
- or Sections 5.3, 5.3, and 8.5 in (Stirzaker 2003).

8.5 Conditional expectation

We now turn to the expectation of a random variable *given* an event (such as the value of another random variable). Recall that the indicator random variable of an event A is given by

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

Then $\mathbb{E}\left[\mathbb{1}\{A\}\right] = 1 \cdot P(A) + 0 \cdot P(A^{c}) = P(A)$.

Key idea: Definition: conditional expectation

Let X be a real-valued random variable, and let $A \subseteq \Omega$ be an event. The *conditional expectation of X given A* is:

$$\mathbb{E}\left[X|A\right] := \frac{\mathbb{E}\left[X\mathbb{1}\{A\}\right]}{P(A)} \text{ whenever } P(A) > 0.$$

Conditional expectation generalizes the concept of conditional probability. For example, if A and B are any events such that $P(B) > 0$ then because $\mathbb{1}\{A\}\mathbb{1}\{B\} = \mathbb{1}\{A \cap B\}$, it follows that $\mathbb{E}\left[\mathbb{1}\{A\} \mid B\right] = \mathbb{E}\left[\mathbb{1}\{A\}\mathbb{1}\{B\}\right] / P(B) = P(A \cap B) \,/ P(B) = P(A \mid B)$.

One may also view $\mathbb{E}\left[\,\cdot\,|A\right]$ as *expectation with respect to the conditional probability* $P(\,\cdot \mid A)$:

Key idea: Theorem: conditional expectation and probabilities

For any discrete random variable X and any event $A \subseteq \Omega$,

$$\mathbb{E}\left[X|A\right] = \sum_{x \in \mathcal{X}} x P(X = x \mid A)\,.$$

Proof

This is an exercise in tracking definitions. Indeed, if $Y = X\mathbb{1}\{A\}$, then for any $x \neq 0$,

$$P(Y = x) = P(\{X = x\} \cap A) = P(X = x \mid A)\, P(A)\,.$$

Hence

$$\begin{aligned} \mathbb{E}[Y] &= \sum_{x \in \mathcal{X}} x P(Y = x) \\ &= \sum_{x \in \mathcal{X},\, x \neq 0} x P(Y = x) \\ &= \sum_{x \in \mathcal{X}} x P(X = x \mid A)\, P(A)\,, \end{aligned}$$

and so

$$\mathbb{E}\left[X|A\right] = \frac{\mathbb{E}[Y]}{P(A)} = \sum_{x \in \mathcal{X}} x P(X = x \mid A)\,,$$

as claimed.

Try it out

Suppose that X and Y are discrete random variables, $g : \mathcal{X} \to \mathbb{R}$, and $B \subseteq \mathcal{Y}$. Then choosing the event $A = \{Y \in B\}$, we see that whenever $P(Y \in B) = \sum_{y \in B} p_Y(y) > 0$,

$$\mathbb{E}\left[g(X)|Y \in B\right] = \frac{\sum_{x \in \mathcal{X}} g(x) \sum_{y \in B} p_{X,Y}(x, y)}{\sum_{y \in B} p_Y(y)}.$$

As a special case, we have

$$\mathbb{E}\left[g(X)|Y = y\right] = \sum_{x \in \mathcal{X}} g(x) p_{X|Y}(x|y).$$

Similarly, if X and Y are jointly continuously distributed, then

$$\mathbb{E}\left[g(X)|Y \in B\right] = \frac{\int_{\mathbb{R}} g(x) \left(\int_B f_{X,Y}(x, y)\, \mathrm{d}y\right) \mathrm{d}x}{\int_B f_Y(y)\, \mathrm{d}y},$$

provided $P(Y \in B) = \int_B f_Y(y)\, \mathrm{d}y > 0$.
To summarize, conditional expectation is just like ordinary expectation but with probabilities replaced by conditional probabilities.

Try it out

In a raffle there is one £500 prize and five £100 prizes. We have one of the 2000 raffle tickets. Let X be our winnings and let A be the event that we have the top prize.

(a) Calculate $\mathbb{E}\left[X \mid A\right]$ and $\mathbb{E}\left[X \mid A^{\mathrm{c}}\right]$.

(b) Compute $\mathbb{E}\left[X|A\right] P(A) + \mathbb{E}\left[X|A^{\mathrm{c}}\right] P(A^{\mathrm{c}})$.

(c) Compute $\mathbb{E}[X]$.

Answer:
By counting we see that $P(X = 500 \mid A) = 1$ and

$$\begin{aligned} P(X = 500 \mid A^{\mathrm{c}}) &= 0, \\ P(X = 100 \mid A^{\mathrm{c}}) &= \frac{5}{1999}, \\ P(X = 0 \mid A^{\mathrm{c}}) &= \frac{1994}{1999}. \end{aligned}$$

So we get $\mathbb{E}\left[X|A\right] = 500$ and

$$\mathbb{E}\left[X|A^{\mathrm{c}}\right] = 100 \cdot \frac{5}{1999} + 0 = \frac{500}{1999}.$$

Now for(b), we note that $P(A) = 1/2000$ and $P(A^{\mathrm{c}}) = 1999/2000$, so

$$\mathbb{E}\left[X|A\right] P(A) + \mathbb{E}\left[X|A^{\mathrm{c}}\right] P(A^{\mathrm{c}}) = 500 \cdot \frac{1}{2000} + \frac{500}{1999} \cdot \frac{1999}{2000} = \frac{1000}{2000} = \frac{1}{2}.$$

For (c), we compute directly that

$$\mathbb{E}[X] = 500 \cdot \frac{1}{2000} + 100 \cdot \frac{5}{2000} + 0 = \frac{1}{2}.$$

So the answers to (b) and (c) are the same.

It was no coincidence that the answers to parts (b) and (c) of were the same. This is an example of the following important theorem.

Key idea: Theorem: partition theorem for expectation

Let X be any real-valued random variable. Let E_1, E_2, ..., E_k be any events that form a partition. Then,

$$\mathbb{E}[X] = \sum_{i=1}^{k} \mathbb{E}\left[X|E_i\right] P(E_i) .$$

Similarly, if E_1, E_2, ... form an infinite partition

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} \mathbb{E}\left[X|E_i\right] P(E_i) .$$

Proof

Because $E_1, E_2, \ldots$ constitute a partition, $\cup_i E_i = \Omega$ and the E_i are pairwise disjoint, so that

$$1 = \mathbb{1}\{\Omega\} = \mathbb{1}\{\cup_i E_i\} = \sum_i \mathbb{1}\{E_i\},$$

and hence, by linearity of expectation,

$$\mathbb{E}[X] = \mathbb{E}\left[X \sum_i \mathbb{1}\{E_i\}\right] = \sum_i \mathbb{E}\left[X \mathbb{1}\{E_i\}\right],$$

which gives the result.

Try it out

Ann and Bob play a sequence of independent games, each with outcomes $A = \{\text{Ann wins}\}$, $B = \{\text{Bob wins}\}$, $D = \{\text{game drawn}\}$ which have probabilities $P(A) = p$, $P(B) = q$ and $P(D) = r > 0$ with $p + q + r = 1$. Let N denote the number of games played until the first win by either Ann or Bob. The results of the first game A, B, D form a partition and $\mathbb{E}\left[N|A\right] = \mathbb{E}\left[N|B\right] = 1$ (as $P(N = 1 \mid A) = P(N = 1 \mid B) = 1$) while $\mathbb{E}\left[N|D\right] = 1 + \mathbb{E}\left[N\right]$ (the future after a drawn game looks the same as at the start). Hence $\mathbb{E}\left[N\right] = 1 \times p + 1 \times q + (1 + \mathbb{E}\left[N\right]) \times r$ or $(1-r)\mathbb{E}\left[N\right] = p + q + r = 1$ i.e. $\mathbb{E}\left[N\right] = 1/(1-r)$.

A more demanding concept is the following:

Key idea: Definition: conditional expectation with respect to a random variable

Let X be a real-valued random variable, and let Y be another random variable. Define the function $g : Y(\Omega) \to \mathbb{R}$ by $g(y) := \mathbb{E}\left[X|Y = y\right]$. Then the *conditional expectation of X given Y* is the random variable denoted by $\mathbb{E}\left[X|Y\right]$ given by

$$\mathbb{E}\left[X|Y\right] := g(Y).$$

You may see this written more compactly in books as

$$\mathbb{E}\left[X|Y\right](\omega) := \mathbb{E}\left[X|Y = Y(\omega)\right], \text{ for all } \omega \in \Omega,$$

but our definition above is a little easier to digest. In the case where Y is discrete, $\mathbb{E}\left[X|Y\right]$ is a random variable that takes values $\mathbb{E}\left[X|Y=y\right]$ with probabilities $P(Y=y)$. We concentrate on the discrete case here.

The next result is of considerable importance.

Key idea: Theorem: partition theorem for expectation II

For any real-valued random variable X and any random variable Y,

$$\mathbb{E}[X] = \mathbb{E}\left[\mathbb{E}\left[X \mid Y\right]\right].$$

Proof

We give a proof in the case when Y is discrete, which shows why we call this a 'Partition theorem'. In this case $\{Y=y\}$, $y \in \mathcal{Y}$, forms a partition, and so gives

$$\mathbb{E}[X] = \sum_{y\in\mathcal{Y}} \mathbb{E}\left[X|Y=y\right] P(Y=y)\,,$$

but this last expression is the expectation of the discrete random variable $\mathbb{E}\left[X|Y\right]$ which takes values $g(y) = \mathbb{E}\left[X|Y=y\right]$ with probabilities $P(Y=y)$:

$$\mathbb{E}\left[\mathbb{E}\left[X \mid Y\right]\right] = \mathbb{E}\left[g(Y)\right] = \sum_{y\in\mathcal{Y}} g(y)P(Y=y)\,,$$

by the law of the unconscious statistician.

This theorem is sometimes called the *law of iterated expectation*. We can use this result to calculate $\mathbb{E}[X]$ in some tricky cases.

Try it out

Toss three fair coins and let H be the total number of heads. The roll a fair die H times, and let T be the total score on the dice rolls. What is $\mathbb{E}\left[T\right]$?
Answer:
This looks tricky, but we use *conditioning* on H to take advantage of the structure of the problem. Note that $H \sim \text{Bin}(3, 1/2)$, so and $\mathbb{E}\left[H\right] = 3/2$. Now, the expected score on a single die is

$$\frac{1}{6}(1+2+3+4+5+6) = \frac{7}{2}.$$

Thus, given $h=h$, the expected total on h rolls of a fair die the expected total score is $\frac{7}{2}h$. In other words, $\mathbb{E}\left[T|H=h\right] = \frac{7}{2}h$. Thus

$$\mathbb{E}\left[T|H\right] = \frac{7}{2}H.$$

Thus, by ,

$$\mathbb{E}\left[T\right] = \mathbb{E}\left[\mathbb{E}\left[T \mid H\right]\right] = \frac{7}{2}\mathbb{E}\left[H\right] = \frac{7}{2} \times \frac{3}{2} = \frac{21}{4}.$$

Note that while this looks very slick, we have actually hidden something here. In fact, we have implicitly used the fact that the scores rolled on the die are independent of H, when we claimed that $\mathbb{E}\left[T|H=h\right] = \frac{7}{2}h$. To see where this is being used, let $S_1, S_2, \ldots$ be the scores on the die rolls, so

$T = \sum_{i=1}^{H} S_i$. Then

$$\mathbb{E}\left[T|H=h\right] = \mathbb{E}\left[\sum_{i=1}^{H} S_i|H=h\right] = \mathbb{E}\left[\sum_{i=1}^{h} S_i|H=h\right],$$

and we can drop the condition here *if* H is independent of the S_i. Then

$$\mathbb{E}\left[T \mid H=h\right] = \mathbb{E}\left[\sum_{i=1}^{h} S_i\right] = \sum_{i=1}^{h} \mathbb{E}\left[S_i\right] = \frac{7}{2}h,$$

as claimed. The next example is of the same type.

Try it out

A shop has N customers a day where $\mathbb{E}\left[N\right] = 800$. Each customer spends £$X_i$ where $\mathbb{E}\left[X_i\right] = 25$. Let $T = \sum_{i=1}^{N} X_i$ be the total takings on a particular day. What is $\mathbb{E}\left[T\right]$?
Answer:
We condition on the value of N. Then

$$\mathbb{E}\left[T|N=n\right] = \mathbb{E}\left[\sum_{i=1}^{n} X_i|N=n\right].$$

This is similar to the last example, and we must here use the fact that the X_i are independent of N to get

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i|N=n\right] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}\left[X_i\right] = 25n.$$

Thus $\mathbb{E}\left[T|N\right] = 25N$ and

$$\mathbb{E}\left[T\right] = \mathbb{E}\left[\mathbb{E}\left[T \mid N\right]\right] = \mathbb{E}\left[25N\right] = 25\mathbb{E}\left[N\right] = 25 \times 800 = 20,000.$$

In this example the assumption of independence between N and the X_i is open to question: if N is very big, perhaps the X_i might be smaller than usual, since the shop runs low on stock, for example.

Textbook references

If you want more help with this section, check out:

- Sections 9.1 and 9.2 in (Blitzstein and Hwang 2019);
- Section 10.3 in (Anderson, Seppäläinen, and Valkó 2018);
- or Sections 4.4 and 5.5 in (Stirzaker 2003).

8.6 Independence: multiplication rule for expectation

Remember that two discrete random variables are independent when their joint probability mass function factorises, i.e. when $p(x,y) = p(x)p(y)$. Similarly, two jointly continuously distributed random variables are independent when their joint probability density function factorizes, i.e. when $f(x,y) = f(x)f(y)$. It turns out that in these cases, expectation factorizes as well:

Key idea: Theorem: Independence means multiply

If X and Y are independent real-valued random variables then

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y].$$

Moreover, if $g, h : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}\left[g(X)h(Y)\right] = \mathbb{E}[g(X)]\mathbb{E}\left[h(Y)\right].$$

More generally, for any mutually independent real-valued random variables X_1, X_2, ..., X_n,

$$\mathbb{E}\left[\prod_{i=1}^{n} X_i\right] = \prod_{i=1}^{n} \mathbb{E}\left[X_i\right].$$

Proof

We give a proof only in the discrete case. Suppose that X and Y are independent discrete random variables with joint probability mass function $p(x,y) = p_X(x)p_Y(y)$. Then by the Law of the Unconscious Statistician,

$$\mathbb{E}\left[g(X)h(Y)\right] = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} g(x)h(y)p(x,y) = \sum_{x\in\mathcal{X}} g(x)p_X(x)\sum_{y\in\mathcal{Y}} h(y)p_Y(y),$$

which is $\mathbb{E}[g(X)]\mathbb{E}\left[h(Y)\right]$.

Corollary: Independence means zero covariance

If X and Y are independent random variables, then $\text{Cov}\left(X, Y\right) = 0$.

Proof

In the independent case, $\text{Cov}\left(X, Y\right) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0$.

The converse of the last property is not true: X and Y may be dependent but uncorrelated.

Example

Suppose that (X, Y) are jointly distributed taking values $(-1, 0)$, $(+1, 0)$, $(0, -1)$, and $(0, +1)$ with probability $1/4$ of each.
Then X and Y are not independent, because $P(X = 1, Y = 1) = 0$ is not the same as $P(X = 1)\,P(Y = 1) = 1/16$, for instance.
However, X and Y are uncorrelated, because $\mathbb{E}[XY] = 0$ and $\mathbb{E}[X] = \mathbb{E}[Y] = 0$ so $\text{Cov}\left(X, Y\right) = 0$. (In fact, in this case $XY = 0$ with probability 1.)

An important consequence of this Corollary is a simplification of the formula for the variance of a sum for pairwise independent random variables:

Corollary: Variance of a sum of independent variables

Consider random variables X_1, X_2, ..., X_n. If these random variables are pairwise independent, then

$$\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \operatorname{Var}\left(X_i\right).$$

Example

Remember from an earlier example that we can write any $X \sim \mathrm{Bin}(n, p)$ as $X = \sum_1^n Y_i$ where Y_1, ..., Y_n are independent and each $Y_i \sim \mathrm{Bin}(1, p)$. In , you showed that $\operatorname{Var}\left(Y_i\right) = p(1-p)$. Consequently, as the Y_i are independent,

$$\operatorname{Var}\left(X\right) = \sum_{i=1}^{n} \operatorname{Var}\left(Y_i\right) = np(1-p).$$

Try it out

Let S be the total score and T be the product of the scores from throwing four dice.
To find $\mathbb{E}\left[S\right]$, $\operatorname{Var}\left(S\right)$ and $\mathbb{E}\left[T\right]$ let the individual scores be X_k, $k = 1, 2, 3, 4$ so that $S = \sum_{k=1}^4 X_k$, $T = \prod_{k=1}^4 X_k$.
We readily calculate $\mathbb{E}\left[X_k\right] = \sum_{i=1}^6 i \times 1/6 = 7/2$ and $\operatorname{Var}\left(X_k\right) = \sum_{i=1}^6 i^2 \times 1/6 - (7/2)^2 = 35/12$. Therefore

$$\mathbb{E}\left[S\right] = \mathbb{E}\left[\sum_{k=1}^{4} X_k\right] = \sum_{k=1}^{4} \mathbb{E}\left[X_k\right] = 4 \times 7/2 = 14,$$

and further, as the X_k are independent,

$$\operatorname{Var}\left(S\right) = \operatorname{Var}\left(\sum_{k=1}^{4} X_k\right) = \sum_{k=1}^{4} \operatorname{Var}\left(X_k\right) = 4 \times 35/12 = 35/3.$$

Again as the X_k are independent,

$$\mathbb{E}\left[T\right] = \mathbb{E}\left[\prod_{k=1}^{4} X_k\right] = \prod_{k=1}^{4} \mathbb{E}\left[X_k\right] = \left(\frac{7}{2}\right)^4 = 2401/16.$$

Note that all of these calculations are possible without having to deal with the joint probability distribution of the X_k (which is uniform on the 1296 possible outcomes).

Textbook references

If you want more help with this section, check out:

- Section7.3 in (Blitzstein and Hwang 2019);
- Section 8.2 in (Anderson, Seppäläinen, and Valkó 2018);
- or Section 5.3 in (Stirzaker 2003).

8.7 Expectation and probability inequalities

By the monotonicity properties of summation and integration, namely that if $f(x) \geq g(x)$ then

$$\sum_i f(x_i) \geq \sum_i g(x_i), \text{ and } \int_A f(x)\,\mathrm{d}x \geq \int_A g(x)\,\mathrm{d}x,$$

we immediately get the following.

Theorem: Monotonicity of expectation

For any random variable X, and any $a \in \mathbb{R}$, if $P(X \geq a) = 1$ then $\mathbb{E}[X] \geq a$.

For instance, suppose that X and Y have $P(X \leq Y) = 1$. Then $P(Y - X \geq 0) = 1$ so $\mathbb{E}[Y - X] \geq 0$ and hence $\mathbb{E}[X] \leq \mathbb{E}[Y]$.

This simple property has various interesting consequences:

Corollary: Variances are positive

For any random variable X, $\mathrm{Var}(X) \geq 0$.

Proof

We have $\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$ and the random variable $(X - \mathbb{E}[X])^2$ is non-negative.

Key idea: Corollary: Markov's inequality

If $X \geq 0$ then, for any $a > 0$,

$$P(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

Proof

Note that $X \geq a\mathbb{1}\{X \geq a\}$, and consequently $0 \leq \mathbb{E}\left[X - a\mathbb{1}\{X \geq a\}\right] = \mathbb{E}[X] - aP(X \geq a)$.

Example

If X equals s with chance p but otherwise equals 0 then $\mathbb{E}[X] = sp$. For $a > s$ we have $0 = P(X \geq a) \leq ps/a$ while for $a \leq s$ Markov's inequality says $p = P(X \geq a) \leq p \times s/a$ which is exact at $a = s$ so this bound is as strong as possible.

Corollary: Chebyshev's inequality

For any random variable X and any $a > 0$, we have

$$P(|X - \mathbb{E}[X]| \geq a) \leq \frac{\mathrm{Var}(X)}{a^2}.$$

Proof

$P((X-\mathbb{E}[X])^2 \geq a^2) \leq \mathbb{E}\left[(X-\mathbb{E}[X])^2\right]/a^2$ by Markov's inequality applied to $(X-\mathbb{E}[X])^2$ at a^2. Now observe that $\{(X-\mathbb{E}[X])^2 \geq a^2\} = \{|X-\mathbb{E}[X]| \geq a\}$ and of course $\mathbb{E}[(X-\mathbb{E}[X])^2] = \text{Var}\,(X)$.

Markov and Chebyshev bounds are often too generous when distributional information is available, as seen in the next examples. Nevertheless, their generality and simplicity make these inequalities very valuable for complex probability calculations.

Try it out

Suppose that $X \sim \text{Bin}(10, 0.1)$. Give an upper bound on $P(X \geq 6)$ using (a) Markov's inequality, and (b) Chebyshev's inequality.

Answer:

For part (a), we get

$$P(X \geq 6) \leq \frac{\mathbb{E}[X]}{6} = \frac{1}{6}.$$

For (b), we get

$$\begin{aligned} P(X \geq 6) &\leq P(|X-1| \geq 5) \\ &= P(|X-\mathbb{E}[X]| \geq 5) \\ &\leq \frac{\text{Var}\,(X)}{25} = \frac{0.9}{25} = 0.036. \end{aligned}$$

The exact probability can be calculated and is $P(X \geq 6) \approx 0.00015$.

Try it out

If $Z \sim \mathcal{N}(0,1)$ then $P(|Z-\mathbb{E}[Z]| \geq 2) = P(|Z| \geq 2) = 1 - P(-2 < Z < 2) = 0.046$, while the Chebyshev bound on this probability is $\text{Var}\,(Z)/2^2 = 0.25$.

Textbook references

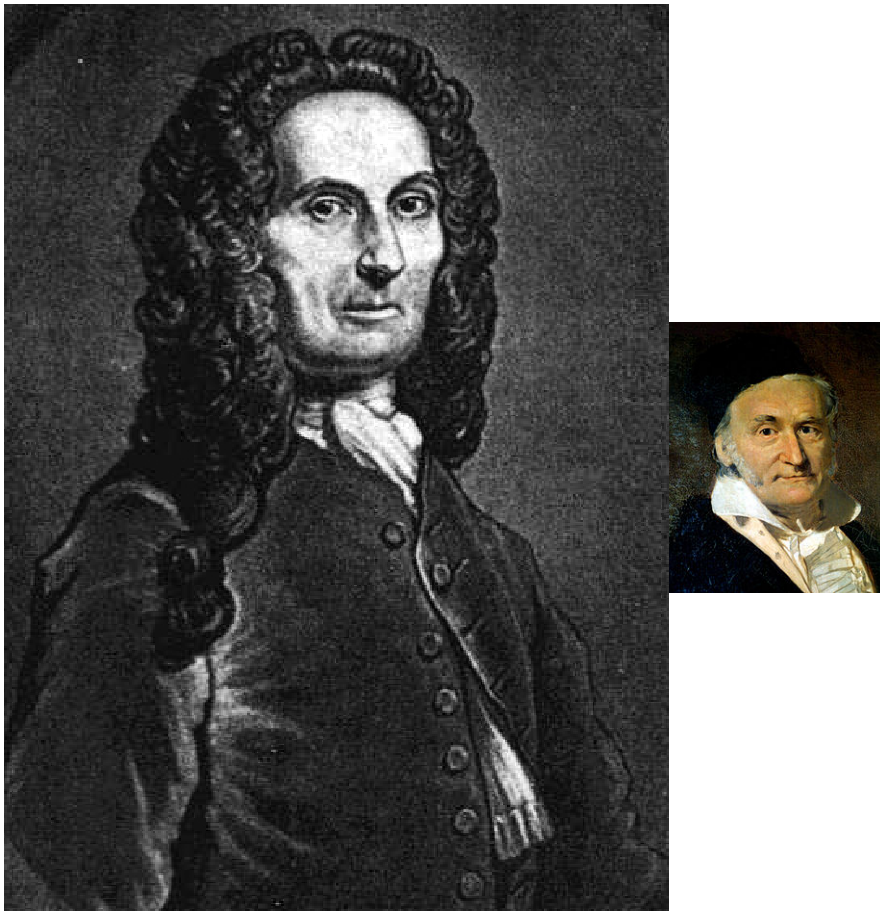
If you want more help with this section, check out:

- Section 10.1 in (Blitzstein and Hwang 2019);
- Section 4. in (DeGroot and Schervish 2013);
- or Section 4.6 in (Stirzaker 2003).

8.8 Historical context

There are approaches to probability theory that start out from expectation of random variables directly, rather than starting out from probability of events as we have done here; see e.g. (Whittle 1992).

Pafnuty Chebyshev (1821–1894) and his student Andrei Markov (1856–1922) made several important contributions to early probability theory. What we call Markov's inequality was actually published by Chebyshev, as was what we call Chebyshev's inequality; our nomenclature is standard, and at least has the benefit of distinguishing the two. A version of the inequality was first formulated by Irénée-Jules Bienaymé (1796–1878).



(b) Markov

(a) Chebyshev

Figure 8.1: (*left to right*) Chebyshev and Markov

9 Limit theorems

Goals

1. Understand and know how to prove the weak law of large numbers, for proportions as well as for general random variables, and know under what conditions the weak law applies.
2. Understand and know how to prove (by means of moment generating functions) the central limit theorem.
3. Know under what conditions the central limit theorem applies.
4. Know how to exploit the central limit theorem to approximate the binomial distribution, and under what circumstances.
5. Know the definition and properties of moment generating functions.
6. Know how to derive the moment generating function of a given distribution.

9.1 The weak law of large numbers

The results of this section describe limiting properties of distributions of sums of random variables using only some assumptions about means and variances.

Toss a coin n times, where the probability of heads is p, independently on each toss. Let X be the number of heads and let $B_n := X/n$ be the proportion of heads in the n tosses. Then, because $X \sim \text{Binom}(n, p)$ has expectation np and variance $np(1-p)$,

$$\mathbb{E}\left[B_n\right] = \mathbb{E}\left[X\right]/n = p, \quad \text{Var}\left(B_n\right) = \text{Var}\left(X\right)/n^2 = p(1-p)/n.$$

So, by Chebyshev's inequality, for any $\epsilon > 0$,

$$P(|B_n - p| \geq \epsilon) \leq \frac{p(1-p)}{n\epsilon^2}.$$

Whence,

$$P(|B_n - p| \geq \epsilon) \to 0 \text{ as } n \to \infty,$$

no matter how small ϵ is. In other words, as $n \to \infty$, the sample proportion is with very high probability within any tiny interval centred on p.

The same argument applies more generally.

Key idea: Theorem: the weak law of large numbers

Suppose we have an infinite sequence X_1, X_2, … of independent random variables with the same mean and variance:

$$\mathbb{E}\left[X_i\right] = \mu \text{ and } \mathrm{Var}\left(X_i\right) = \sigma^2 \text{ for all } i.$$

Consider the sample average $\bar{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$. Then, for any $\epsilon > 0$,

$$\lim_{n\to\infty} P\big(|\bar{X}_n - \mu| > \epsilon\big) = 0. \tag{9.1}$$

In other words, the sample average has a very high probability of being very near the expected value μ when n is large. The type of convergence in Equation 9.1 is called *convergence in probability*: the weak law of large numbers says that "$\bar{X}_n$ converges in probability to μ".

Proof

We use Chebyshev's inequality to bound the probability that we are trying to show is small. First note that, by linearity of expectation, $\mathbb{E}\left[\bar{X}_n\right] = \frac{1}{n}\sum_{i=1}\mathbb{E}\left[X_i\right] = \mu$ and, by independence, $\mathrm{Var}\left(\bar{X}_n\right) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}\left(X_i\right) = \frac{\sigma^2}{n}$. So, by Chebyshev's inequality,

$$P(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2},$$

which indeed converges to zero as n tends to infinity.

Advanced content

The assumption of finite variances in this theorem is not necessary. The weak law of large numbers holds assuming only that $\mathbb{E}\left[X_i\right] = \mu$ for independent X_i: for this and other more advanced results, see (Feller 1968, chap. 10).

Examples

1. Measure the heights H_i of n randomly selected people from a very large population, where μ, σ^2 are the average and the variance of heights over the whole population.

   Then $\mathbb{E}\left[H_i\right] = \mu$, $\mathrm{Var}\left(H_i\right) = \sigma^2$ and so, as long as the collection of heights is not too asymmetric, the chance of the average height $\bar{X}_n$ being more than a small amount from μ is very small; i.e. almost all large samples have average near μ.

2. Here is an application to repeated sampling. Let X be a random variable whose distribution we want to study by 'sampling', i.e., observing a number of independent random variables $X_1, X_2, \dots, X_n$ which all have the same distribution as X. We could observe $X_1, X_2, \dots, X_n$ and consider

   $$\pi_n(a,b) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{X_i \in [a,b)\},$$

   the proportion of observations whose value falls in the interval $[a,b)$. Since the X_i are independent, so are the indicator random variables, and we know that $\mathbb{E}\left[\mathbb{1}\{X_i \in [a,b)\}\right] = P(X_i \in [a,b))$. So the law of large numbers says that $\pi_n(a,b)$ will approach $P(X_i \in [a,b))$ for large n with high probability. Another way to see this is to observe that $\sum_{i=1}^n \mathbb{1}\{X_i \in [a,b)\}$ is a binomial random variable.

If we want to look at the distribution of X, we would construct a histogram using the proportions $\pi_n(a_i, b_i)$ over a collection of 'bins' $[a_i, b_i)$. If there are only finitely many bins, then it follows that all the π_n's tend to be close to their respective probabilities. For example, the picture in Figure 9.1 shows a histogram produced by 10^4 simulations of a $U(0, 1)$ random variable: the fact that the histogram is a good approximation to the probability density function can, in this instance, be seen as a consequence of the law of large numbers.

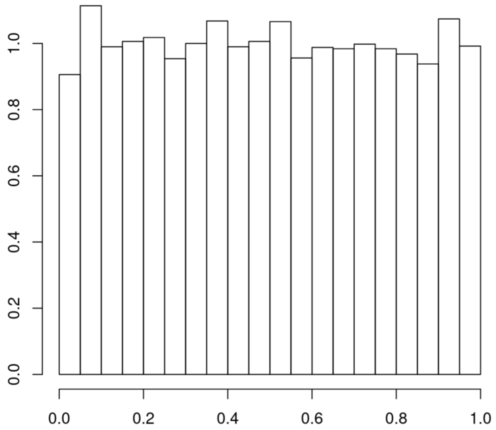


Figure 9.1: Histogram generated from 10^4 simulations of a $\mathrm{U}(0, 1)$ random variable. The vertical axis is the frequency.

Textbook references

If you want more help with this section, check out:

- Section 10.2 in (Blitzstein and Hwang 2019);
- Section 9.2 in (Anderson, Seppäläinen, and Valkó 2018);
- or Section 5.8 in (Stirzaker 2003).

9.2 The central limit theorem

If the law of large numbers is a 'first order' result, a 'second order result' is the famous central limit theorem, which describes *fluctuations* around the law of large numbers, and explains, in part, why the normal distribution has a central role in statistics. A sequence of random variables $X_1, X_2, \dots$ are *independent and identically distributed* (*i.i.d.* for short) if they are mutually independent and all have the same (marginal) distribution.

Key idea: Theorem: the Central Limit Theorem

Suppose we have a sequence X_1, X_2, ...of i.i.d. random variables. Let

$$\mu := \mathbb{E}\left[X_i\right] \text{ and } \sigma^2 \quad := \mathrm{Var}\left(X_i\right),$$

with $\sigma > 0$. Let $S_n := \sum_{i=1}^n X_i$, $\bar{X}_n := S_n/n$, and

$$Z_n := \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

Then, for any $z \in \mathbb{R}$,

$$\lim_{n\to\infty} F_{Z_n}(z) = \lim_{n\to\infty} P(Z_n \leq z) = \Phi(z).$$

We say that Z_n *converges in distribution* to the standard normal distribution.

In other words, for large n, we have that approximately $Z_n \approx \mathcal{N}(0,1)$. Note that the definition of Z_n is such that $\mathbb{E}[Z_n] = 0$ and $\mathrm{Var}(Z_n) = 1$. The content of the central limit theorem is that it should be approximately normal. Consequently, if we also invoke , we approximately have that $S_n \approx \mathcal{N}(n\mu, n\sigma^2)$ and $\bar{X}_n \approx \mathcal{N}(\mu, \sigma^2/n)$, i.e., typical values of S_n are of order $\sigma\sqrt{n}$ from $n\mu$ while typical values of $\bar{X}_n$ are of order $\sigma/\sqrt{n}$ from μ. In other words, for large enough n, by ,

$$P(S_n \leq x) \approx \Phi\left(\frac{x - n\mu}{\sigma\sqrt{n}}\right),$$
$$P(\bar{X}_n \leq x) \approx \Phi\left(\frac{x - \mu}{\sigma/\sqrt{n}}\right).$$

Example

Suppose $X_1, X_2, \ldots$ are i.i.d. exponential random variables with parameter 1, i.e., they have probability density function $f(x) = e^{-x}$ for $x > 0$ and $f(x) = 0$ otherwise.

Consider $S_n = \sum_{i=1}^n X_i$. The central limit theorem says that S_n will be approximately normal for large n, and since we know $\mathbb{E}[X_i] = 1$ and $\mathrm{Var}(X_i) = 1$ in this case, S_n will be approximately $\mathcal{N}(n,n)$ for large n.

How "large" should n be? Well, in this case it turns out we can compute the distribution of S_n exactly. It is an example of a *gamma distribution*, and for $n \geq 1$, S_n has probability density function

$$f_n(x) = \begin{cases} \dfrac{e^{-x}x^{n-1}}{(n-1)!} & \text{if } x > 0, \\ 0 & \text{elsewhere.} \end{cases}$$

See Figure 9.2 for some plots of this for various values of n.

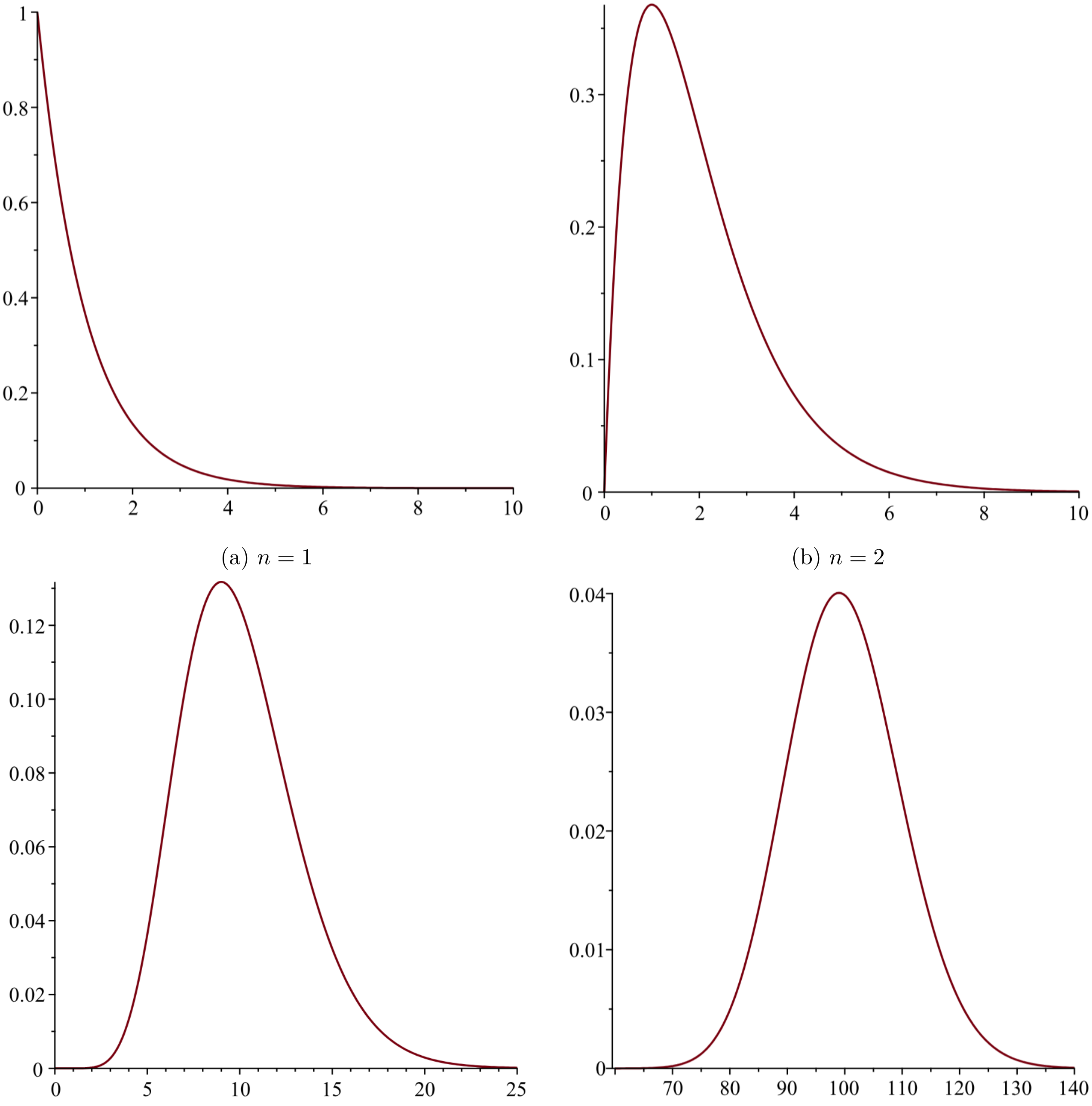
Advanced content

The conditions of the central limit theorem can be substantially weakened.

The assumption that the X_i are identically distributed can be dropped and replaced with the *Lindeberg condition*: see e.g. (Feller 1971, chap. VIII).

In fact, the random variables X_1, X_2,... need be neither identical nor independent: it is sufficient that we can turn them into a *normalized martingale*. Without going into too much detail, it suffices to assume that

$$\mathbb{E}[X_{n+1} | X_1 = x_1 \ldots X_n = x_n] = \mathbb{E}[X_{n+1}]$$
$$\mathrm{Var}(X_{n+1} | X_1 = x_1 \ldots X_n = x_n) = \mathrm{Var}(X_{n+1}) > 0$$



(a) $n = 1$

(b) $n = 2$

(c) $n = 10$

(d) $n = 100$

Figure 9.2: Plots of f_n for $n = 1, 2$ (top row) and $n = 10, 100$ (bottom row).

for all n and all possible values for x_1, ...,x_n. In this case, the sequence of random variables

$$Z_n := \frac{\sum_{k=1}^{n}(X_k - \mathbb{E}[X_k])}{\sqrt{\sum_{k=1}^{n} \operatorname{Var}(X_k)}}$$

converges to a $\mathcal{N}(0,1)$ random variable whenever the martingale version of the Lindeberg condition is satisfied. For further details, see for instance (Nelson 1987, chap. 14 & 18).

Try it out

Potatoes with an average weight of 100g and standard deviation of 40g are packed into bags to contain at least 2500g. What is the chance that more than 30 potatoes will be needed to fill a given bag?

Answer:

Let N be the number needed to exceed 2500 and W be the total weight of 30 potatoes, $W = \sum_{i=1}^{30} X_i$. Then $\mathbb{E}[W] = 30 \cdot 100 = 3000$ and $\operatorname{Var}(W) = 30 \cdot 40^2 = 48000$, and the central limit theorem says that

$$\frac{W - 3000}{\sqrt{48000}} \approx \mathcal{N}(0,1).$$

Also, $\{N > 30\} = \{W < 2500\}$ and so

$$\begin{aligned} P(N > 30) &= P(W < 2500) \\ &= P\left(\frac{W - 3000}{\sqrt{48000}} < \frac{2500 - 3000}{\sqrt{48000}}\right) \\ &\approx P(Z < -2.282), \end{aligned}$$

where $Z \sim \mathcal{N}(0,1)$. From the tables, this probability is

$$P(Z < -2.282) = P(Z > 2.282) = 1 - \Phi(2.282) \approx 0.011.$$

Try it out

Measurements from a particular experiment have mean μ (unknown) and known standard deviation $\sigma = 2.5$. We perform 20 repetitions of the experiment and use $\bar{X}_{20}$ to estimate μ. What is $P\big(|\bar{X}_{20} - \mu| < 1\big)$?

Answer:

The central limit theorem says that

$$\frac{\bar{X}_n - \mu}{\sqrt{2.5^2/n}} \approx \mathcal{N}(0,1).$$

So

$$\begin{aligned} P\big(|\bar{X}_{20} - \mu| < 1\big) &= P\big(-1 \leq \bar{X}_{20} - \mu \leq 1\big) \\ &= P\left(-\frac{1}{\sqrt{2.5^2/20}} \leq \frac{\bar{X}_{20} - \mu}{\sqrt{2.5^2/20}} \leq \frac{1}{\sqrt{2.5^2/20}}\right) \\ &\approx P(-1.789 \leq Z \leq 1.789), \end{aligned}$$

where $Z \sim \mathcal{N}(0,1)$. Thus

$$\begin{aligned} P\big(|\bar{X}_{20} - \mu| < 1\big) &\approx \Phi(1.789) - \Phi(-1.789) \\ &= 2\Phi(1.789) - 1 \\ &\approx 0.92, \end{aligned}$$

using normal tables.
Note that Chebyshev's inequality gives the weaker result

$$\begin{aligned} P\big(|\bar{X}_{20} - \mu| < 1\big) &= 1 - P\big(|\bar{X}_{20} - \mu| \geq 1\big) \\ &\geq 1 - \frac{\mathrm{Var}\left(\bar{X}_{20}\right)}{1^2} \\ &= 1 - \frac{2.5^2}{20} \approx 0.69. \end{aligned}$$

Remember that we can write any binomially distributed random variable as a sum of independent Bernoulli random variables. Consequently:

Corollary: Normal approximation to the Binomial

Let $X \sim \mathrm{Bin}(n,p)$ with $0 < p < 1$, $B_n := X/n$, and

$$Z_n := \frac{X - np}{\sqrt{np(1-p)}} = \frac{B_n - p}{\sqrt{p(1-p)/n}}$$

Then

$$\lim_{n\to\infty} F_{Z_n}(z) = \lim_{n\to\infty} P(Z_n \leq z) = \Phi(z).$$

This means that, when $X \sim \mathrm{Bin}(n,p)$ with $0 < p < 1$ and large enough n, then for any $x \in \mathbb{R}$, we approximately have that

$$P(X \leq x) \approx \Phi\left(\frac{x - np}{\sqrt{np(1-p)}}\right)$$

For moderate n, a continuity correction improves the approximation; e.g. for $k \in \mathbb{N}$:

$$\begin{aligned} P(X \leq k) &= P(X \leq k + 0.5) \approx \Phi\left(\frac{k + 0.5 - \mu}{\sigma}\right) \\ P(k \leq X) &= P(k - 0.5 \leq X) \approx 1 - \Phi\left(\frac{k - 0.5 - \mu}{\sigma}\right) \end{aligned}$$

Try it out

Consider a multiple choice test with 50 questions, one mark for each correct answer. Independently for each question, a particular student has chance 1/2 of answering correctly. Find, approximately, the probability of the student scoring at least 30 marks.

Answer

Let X = student's score, so $X \sim \mathrm{Bin}(50, 1/2)$. Then $\mathbb{E}[X] = 50/2 = 25$ and $\mathrm{Var}(X) = 50/4 = 25/2$.

The normal approximation says that $X \approx \mathcal{N}(25, 25/2)$, so

$$\begin{aligned} P(X \geq 30) &= 1 - P(X < 30) \\ &\approx 1 - \Phi\left(\frac{30-25}{3.54}\right) = 1 - \Phi(1.41) \approx 0.1. \end{aligned}$$

Try it out

A plane has 110 seats and n business people book seats. They show up independently for their flight with chance $q = 0.85$. Let X_n be the number that arrive to take the flight (the others just take a different flight) and find $P(X_n > 110)$ for $n = 110, 120, 130, 140$.

Answer:

Using the CLT for binomials, $\mathbb{E}[X_n] = 0.85n$ and $\sigma(X_n) = \sqrt{nq(1-q)} = 0.3571\sqrt{n}$ and so $P(X_n > 110) \approx 1 - \Phi((110.5 - 0.85n)/0.3571\sqrt{n})$ which takes values 0, 0.015, 0.500 and 0.978 for $n = 110, 120, 130$ and 140.

The proof of the central limit theorem requires an important new tool: the moment generating function.

Textbook references

If you want more help with this section, check out:

- Section 10.3 in (Blitzstein and Hwang 2019);
- Section 9.3 in (Anderson, Seppäläinen, and Valkó 2018);
- or Section 8.9 in (Stirzaker 2003).

9.3 Moment generating functions

Key idea: Definition: moment generating function

For any real-valued random variable X, the function $M_X : \mathbb{R} \to [0, +\infty]$ given by

$$M_X(t) := \mathbb{E}\left[e^{tX}\right]$$

is called the *moment generating function* of X.

Because $e^{tX} \geq 0$, we have that $M_X(t) \geq 0$ by monotonicity of expectations.

Using the Law of the Unconscious Statistician, we can derive the following expressions for $M_X(t)$:

$$\begin{aligned} M_X(t) &= \sum_{x \in \mathcal{X}} e^{tx} p(x) && \text{if } X \text{ is discrete, and} \\ M_X(t) &= \int_{-\infty}^{\infty} e^{tx} f(x) \mathrm{d}x && \text{if } X \text{ is continuously distributed.} \end{aligned}$$

The above sum and integral always exist, but can be $+\infty$.

Examples

1. If $X \sim \text{Bin}(1, p)$ then $M_X(t) = pe^t + (1-p)$.
2. If $Y \sim \text{Po}(\lambda)$ then
$$\begin{aligned} M_Y(t) &= \sum_{x=0}^{\infty} e^{tx} p(x) \\ &= \sum_{x=0}^{\infty} e^{-\lambda} \frac{\lambda^x}{x!} e^{tx} \\ &= \sum_{x=0}^{\infty} e^{-\lambda} \frac{(\lambda e^t)^x}{x!} \\ &= \exp(\lambda(e^t - 1)). \end{aligned}$$
3. If $U \sim \text{U}(a, b)$ then
$$M_U(t) = \frac{e^{bt} - e^{at}}{(b-a)t}$$
for $t \neq 0$, and $M_U(0) = 1$.
4. If $Z \sim \mathcal{N}(0, 1)$ then
$$\begin{aligned} M_Z(t) &= \int_{-\infty}^{\infty} f_Z(z) e^{tz} \mathrm{d}z \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} e^{tz} \mathrm{d}z \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(z-t)^2/2} e^{t^2/2} \mathrm{d}z, \end{aligned}$$
as we see by completing the square in the exponential. Now put $y = z - t$ to get
$$M_Z(t) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} e^{t^2/2} \mathrm{d}y = e^{t^2/2},$$
because the y-dependent part is just $f_Z(y)$, which integrates to 1.

The moment generating function has several useful properties. The property that gives the name is revealed by considering the Taylor series for e^{tX}: formally,

$$\begin{aligned} M_X(t) = \mathbb{E}\left[e^{tX}\right] &= \mathbb{E}\left[1 + tX + \frac{t^2 X^2}{2!} + \frac{t^3 X^3}{3!} + \cdots\right] \\ &= 1 + t\mathbb{E}\left[X\right] + \frac{t^2}{2!}\mathbb{E}\left[X^2\right] + \frac{t^3}{3!}\mathbb{E}\left[X^3\right] + \cdots, \end{aligned}$$

at least if $t \approx 0$. (Some work is needed to justify this last step, which we omit.) This gives the first of our properties.

Key idea: Properties of moment generating functions

M1: *(Moment generating functions generate moments.)*
For every $k \in \mathbb{N}$,
$$\mathbb{E}\left[X^k\right] = \frac{d^k M_X}{dt^k}(0).$$

M2: *(Moment generating function determines distribution.)*
Consider any two random variables X and Y. If there is an $h > 0$ such that

$$M_X(t) = M_Y(t) < +\infty \qquad \text{for all } t \in (-h, h),$$

then

$$F_X(x) = F_Y(x) \qquad \text{for all } x \in \mathbb{R}.$$

Conversely, if $F_X(x) = F_Y(x)$ for all $x \in \mathbb{R}$ then $M_X(t) = M_Y(t)$ for all $t \in \mathbb{R}$.
M3: *(Scaling.)*
For any random variable X and any constants a, $b \in \mathbb{R}$,

$$M_{aX+b}(t) = e^{bt} M_X(at).$$

M4: *(Product.)*
Suppose that X_1, ..., X_n are independent random variables and let $Y = \sum_{i=1}^n X_i$. Then

$$M_Y(t) = \prod_{i=1}^n M_{X_i}(t).$$

M5: *(Convergence.)*
Suppose that X_1, X_2, ...is an infinite sequence of random variables, and that X is a further random variable. If there is an $h > 0$ such that

$$\lim_{n\to\infty} M_{X_n}(t) = M_X(t) < +\infty \qquad \text{for all } t \in (-h, h),$$

then

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x) \qquad \text{for all } x \in \mathbb{R} \text{ where } F_X \text{ is continuous},$$

i.e., X_n converges in distribution to X.

Proof

The proof of **M3** just uses linearity of expectation.
M1, **M2**, and **M5** use some deeper analysis, which we omit.
For **M4**, we use the fact that “independence means multiply” when working with expectations:

$$\begin{aligned}
M_{S_n}(t) &= \mathbb{E}\left[e^{t(X_1+X_2+\cdots+X_n)}\right] \\
&= \mathbb{E}\left[e^{tX_1} e^{tX_2} \cdots e^{tX_n}\right] \\
&= \mathbb{E}\left[e^{tX_1}\right] \mathbb{E}\left[e^{tX_2}\right] \cdots \mathbb{E}\left[e^{tX_n}\right] \\
&= \prod_{i=1}^n M_{X_i}(t).
\end{aligned}$$

Advanced content

Regarding M5 (convergence), in the case of the central limit theorem, the limit has $F_X(x) = \Phi(x)$ which is continuous for all $x \in \mathbb{R}$, so in that case, F_{X_n} converges to F_X everywhere. As we saw earlier (), F_X is in fact continuous whenever X is continuously distributed, so in that case convergence to

$F_X(x)$ takes place for all x.
In general, one can show that the cumulative distribution function F of *any* random variable X has at most a countable number of points where F is not continuous.
To see this, let

$$D_n = \left\{ x \in \mathbb{R} : F(x) - F(x-) > \frac{1}{n} \right\},$$

the points x at which $F(x)$ has a jump of size at least $1/n$. Then the points at which F is discontinuous can be expressed as

$$D = \bigcup_{n \in \mathbb{N}} D_n.$$

But D_n is a finite set since F is non-decreasing; indeed, D_n is at most of size n, or else the jumps would add up to more than 1, which is impossible. So D is a countable union of finite sets, and is hence countable.

Examples

1. Suppose that $Z \sim \mathcal{N}(0,1)$. We saw earlier that $M_Z(t) = e^{t^2/2}$.

   So, by **M1**,

   $$\mathbb{E}\left[Z\right] = M_Z'(0) = te^{t^2/2}\Big|_{t=0} = 0,$$

   and

   $$\mathbb{E}\left[Z^2\right] = M_Z''(0) = (1+t^2)e^{t^2/2}\Big|_{t=0} = 1,$$

   so $\mathrm{Var}\left(Z\right) = \mathbb{E}\left[Z^2\right] - \mathbb{E}\left[Z\right]^2 = 1 - 0 = 1$.

2. Suppose $X \sim \mathrm{Po}(\lambda)$. Then, by **M1**,

   $$\mathbb{E}\left[X\right] = M_X'(0) = \lambda e^0 \exp(\lambda(e^0 - 1)) = \lambda.$$

3. If $Z \sim \mathcal{N}(0,1)$, then by **M3**,

   $$M_{\mu+\sigma Z}(t) = e^{\mu t} M_Z(\sigma t) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right).$$

   We already know that $\mu + \sigma Z \sim \mathcal{N}(\mu, \sigma^2)$ (see the "standardising the normal distribution" theorem from Chapter 6), so the above expression gives us the moment generating function of the normal distribution with mean μ and variance σ^2.

   Additionally, by **M2**, it follows that $X \sim \mathcal{N}(\mu, \sigma^2)$ if and only if $M_X(t) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)$.

4. Suppose that X_1, ..., X_k are independent random variables with $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$.

   If $Y = \sum_{i=1}^k X_i$, then, by M4, the moment generating function of Y is

   $$M_Y(t) = \prod_{i=1}^k \exp\left(\mu_i t + \frac{1}{2}\sigma_i^2 t^2\right) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)$$

   where $\mu = \sum_{i=1}^k \mu_i$ and $\sigma^2 = \sum_{i=1}^k \sigma_i^2$. Thus, by **M2**, it must be that $Y \sim \mathcal{N}(\mu, \sigma^2)$.

Try it out

Let $X_1, \dots, X_n$ be independent with $X_i \sim \text{Po}(\lambda_i)$. Identify the distribution of $Y_n = \sum_{i=1}^n X_i$.
Answer:
By **M4** and our earlier calculation of the mgf of a Possion random variable,

$$\begin{aligned} M_Y(t) &= \prod_{i=1}^n M_{X_i}(t) \\ &= \prod_{i=1}^n \exp\left(\lambda_i(e^t - 1)\right) \\ &= \exp\left(\left(\sum_{i=1}^n \lambda_i\right)(e^t-1)\right). \end{aligned}$$

By uniqueness (**M2**) this is the moment generating function of $\text{Po}(\sum_{i=1}^n \lambda_i)$. So a sum of independent Poissons is also Poisson!

Proof: Proof of the Central Limit Theorem

Recall from calculus: if $a_n \to a$ then

$$\left(1 + \frac{a_n}{n}\right)^n \to e^a.$$

We want to show that for

$$Z_n = \sum_{i=1}^n \frac{X_i - \mu}{\sigma\sqrt{n}}$$

we have $M_{Z_n}(t) \to e^{t^2/2}$ for all t in an open interval containing 0. This will give the central limit theorem by M5.
Let $Y_i = (X_i - \mu)/\sigma$ and denote the moment generating function of the Y_i by $m(t)$. By M3, $Y_i/\sqrt{n}$ has moment generating function $m(t/\sqrt{n})$. Then by M4, $Z_n = \sum_{i=1}^n Y_i/\sqrt{n}$ has moment generating function

$$M_{Z_n}(t) = \left(m(t/\sqrt{n})\right)^n.$$

Next, by M1,

$$m(0) = \mathbb{E}\left[Y_i^0\right] = 1, \quad m'(0) = \mathbb{E}\left[Y_i\right] = 0, \quad m''(0) = \mathbb{E}\left[Y_i^2\right] = 1.$$

So by Taylor's theorem around 0 there is a function h with $h(u) \to 0$ such that

$$m(u) = 1 + \frac{u^2}{2} + u^2 h(u).$$

Hence

$$M_{Z_n}(t) = \left(1 + \frac{t^2}{2n} + \frac{t^2}{n} h(t/\sqrt{n})\right)^n \to e^{t^2/2},$$

for any $t \in \mathbb{R}$.

Textbook references

For more help with this section, check out:

- Section 6.4 in(Blitzstein and Hwang 2019);

- Section 5.1 in (Anderson, Seppäläinen, and Valkó 2018);
- or Sections 6.4 and 7.5 in (Stirzaker 2003).

9.4 Historical context

The law of large numbers and the central limit theorem have long and interesting histories. The weak law of large numbers for binomial (i.e. sums of Bernoulli) variables was first established by Jacob Bernoulli (1654–1705) and published in 1713 (Bernoulli 1713). The name 'law of large numbers' was given by Poisson. The modern version is due to Aleksandr Khinchin (1894–1959), and our Central Limit Theorem is only a special case—the assumption on variances is unnecessary.

(a) Bernoulli

(b) Khinchin

(c) Lyapunov

(d) Polya

It was apparent to mathematicians in the mid 1700s that a more refined result than Bernoulli's law of large numbers could be obtained. A special case of the central limit theorem for binomial (i.e. sums of Bernoulli) variables was first established by de Moivre in 1733, and extended by Laplace; hence the normal approximation to the binomial is sometimes known as the *de Moivre–Laplace theorem*. The name 'central limit theorem' was given by George Pólya (1887–1985) in 1920.

The first modern proof of the central limit theorem was given by Aleksandr Lyapunov (1857–1918) around 1901 ("Lyapunov Theorem," n.d.). Lyapunov's assumptions were relaxed by Jarl Waldemar Lindeberg (1876–1932) in 1922 (Lindeberg 1922). Many different versions of the central limit theorem were subsequently proved. The subject of Alan Turing's (1912–1954) Cambridge University Fellowship Dissertation of 1934 was a version of the central limit theorem similar to Lindeberg's; Turing was unaware of the latter's work.

References

Anderson, D. F., T. Seppäläinen, and B. Valkó. 2018. *Introduction to Probability.* Cambridge University Press.

Bernoulli, J. 1713. *Ars Conjectandi.* Basileæ: Thurnisiorum.

Billinton, R., and R. N. Allan. 1996. *Reliability Evaluation of Power Systems.* 2nd ed. Plenum Press.

Blitzstein, J. K., and J. Hwang. 2019. *Introduction to Probability.* Texts in Statistical Science Series. CRC Press.

Boole, G. 1854. *An Investigation of the Laws of Thought: On Which Are Founded the Mathematical Theories of Logic and Probabilities.* London: Walton; Maberly.

Chung, K. L., and F. AitSahlia. 2003. *Elementary Probability Theory.* 4th ed. Undergraduate Texts in Mathematics. Springer-Verlag, New York.

DeGroot, M. H., and M. J. Schervish. 2013. *Probability and Statistics.* 4th ed. Harlow, England: Pearson.

Feller, W. 1968. *An Introduction to Probability Theory and Its Applications, Vol. 1.* 3rd ed. Wiley, New York.

———. 1971. *An Introduction to Probability Theory and Its Applications, Vol. 2.* 2nd ed. Wiley, New York.

Hacking, I. 2006. *The Emergence of Probability.* Cambridge University Press.

Hájek, A. 2012. "Interpretations of Probability." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta.

Kolmogorov, A. N. 1950. *Foundations of the Theory of Probability.* New York: Chelsea Publishing Company.

Laplace, P. S. 1825. *Essai Philosophique Sur Les Probabilitiés.* Paris: Bachelier.

Lindeberg, J. W. 1922. "Eine Neue Herleitung Des Exponentialgesetzes in Der Wahrscheinlichkeitsrechnung." *Mathematische Zeitschrift* 15: 211–25.

"Lyapunov Theorem." n.d. Encyclopedia of Mathematics.

Mahmoud, H. M. 2009. *Pólya Urn Models.* CRC Press, Boca Raton, FL.

Moivre, A. de. 1756. *The Doctrine of Chances: Or, a Method for Calculating the Probabilities of Events in Play.* Third. London: A. Millar.

Moran, P. A. P. 1968. *An Introduction to Probability Theory.* Clarendon Press, Oxford.

Nelson, E. 1987. *Radically Elementary Probability Theory.* Princeton University Press.

Rosenthal, J. 2007. *A First Look at Rigorous Probability Theory.* Second. New York: World Scientific.

Ross, S. M. 2010. *Introduction to Probability Models.* 10th ed. Academic Press, Amsterdam.

Stirzaker, D. 2003. *Elementary Probability.* Second. Cambridge University Press.

Todhunter, I. 2014. *A History of the Mathematical Theory of Probability.* Cambridge University Press.

Venn, J. 1888. *The Logic of Chance: An Essay on the Foundations and Province of the Theory of Probability, with Especial Reference to Its Application to Moral and Social Science.* Third. London: Macmillan.

Whittle, P. 1992. *Probability via Expectation.* Third. New York: Springer.

Whitworth, W. A. 1901. *Choice and Chance.* Third. Cambridge: Deighton Bell.