Chris Hallsworth

Probability for Statistics

October 5, 2021

Contents

1 *Probability Review* 5

2 *Random Variables* 17

3 *Multivariate Random Variables* 37

4 *Convergence of Random Variables* 51

5 *Central Limit Theorem* 63

6 *Stochastic Processes* 71

1

Probability Review

We will begin by developing a framework to reason about the possible outcomes of real-world experiments. Our framework will be an abstraction of the process of collecting data from a system that is subject to chance variation. The classical examples are flipping a coin, or rolling a die. Practically relevant examples are clinical trials, experiments on biological systems, physical measurements with limited precision, or observations of complex systems such as the financial markets or social networks. The important common factor is that in any realistic setting, the input and surroundings may vary in ways we cannot perceive, let alone control. We therefore allow explicitly for apparently identical inputs to produce different outputs.

It will be important to distinguish betweeen *what can happen* and *what we can observe*. To make this distinction, we define the sample space, which is the set of possible outcomes of our experiment. Each observable event will correspond to a subset of the sample space. The collection of all such events will satisfy some natural closure properties, which correspond to the logical consequences of our observations. Probabilities, when we define them, will only be assigned to these observable events.

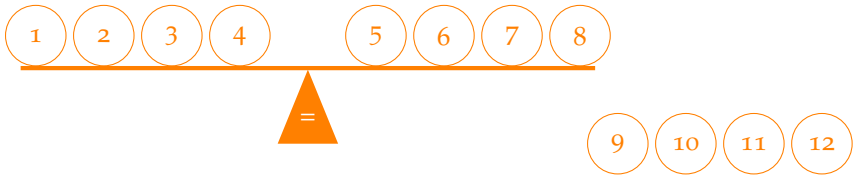


Figure 1.1: A two-pan balance for use with the example 1.3

Definition 1.1. *An* **experiment** *is any fixed procedure with a variable outcome.*

Definition 1.2. *The* **sample space** *is the set of possible outcomes of an*

experiment. We denote the sample space by Ω.

Example 1.3. *We'll use the set-up of a well-known logic puzzle to illustrate what we mean by an experiment, an outcome and an event. Suppose we have 12 balls of identical appearance. 11 of the balls have identical mass, and one of them has a different mass. Suppose we are given a balance, as shown in Figure 1.1. The aim of the original puzzle is to determine which ball is different, and whether it is heavier or lighter than the others, using the balance as few times as possible. If you have not seen it before, it is a very nice problem to think about.*

In this example, we focus just on which ball is different, so let's suppose we know that the different ball is lighter than the others. There are 12 different possibilities, which we might regard as states of the world. These are the **outcomes** *of our* **experiment**. *The* **sample space** *is the set of all outcomes,*

$$\Omega = \{1, 2, \ldots, 12\}.$$

Suppose we use the balance to compare the total mass of balls $1, 2, 3, 4$ *with that of balls* $5, 6, 7, 8$. *What information is available to us about* $\omega \in \Omega$, *the ball that is different, from this single use of the balance? What are the* **events** *that we might observe?*
We could make three different observations. In the obvious notation, these are

$$\{1,2,3,4\} < \{5,6,7,8\}, \qquad \{1,2,3,4\} = \{5,6,7,8\}, \qquad \{1,2,3,4\} > \{5,6,7,8\},$$

These observations correpond, respectively, to the following findings about ω.

$$\omega \in A = \{1,2,3,4\}, \qquad \omega \in B = \{9,10,11,12\}, \qquad \omega \in C = \{5,6,7,8\}.$$

After this weighing, we can state definitively whether or not $\omega \in A$, *for certain subsets* $E \subseteq \Omega$. *These are the* **events**, *the subsets about which we can draw conclusions. Logically, the collection of all events must obey certain closure properties, namely, those for an* **algebra** *of sets :*

1. *We always know that* $\omega \notin \emptyset$, *or equivalently we know* $\omega \in \Omega$.
2. *If we know whether or not* $\omega \in E$, *then we know whether or not* $\omega \in E^c = \Omega \backslash E$.
3. *If we know whether or not* $\omega \in E$ *and whether or not* $\omega \in F$, *then we must know whether or not* $\omega \in E \cup F$.

For the experiment described, the collection of events is

$$\mathcal{F}_0 = \{\emptyset,\ A,\ B,\ C,\ A \cup B,\ A \cup C,\ B \cup C,\ \Omega\}.$$

Of course, if we were allowed to use the balance a few more times, we would be able to determine once and for all which of the balls is different, and moreover, we could design a procedure that achieved this aim no matter which $\omega \in \Omega$ occurred. Then, the collection of observable events would be the power set of Ω - the set of all subsets of Ω.

For discrete problems, in general, i.e. those with a finite or countably infinite state space, in theory there is nothing to stop us taking the sigma algebra of events to be the power set of the sample space. This just corresponds to having sufficient resolution to distinguish which outcome has occurred. Nonetheless, the conceptual distinction between the outcome of an experiment $\omega \in \Omega$ - what can happen, and an event $E \subseteq \Omega$, with $E \in \mathcal{F}$ - what we can observe or measure, is an important one when answering practical questions in statistics.

Definition 1.4. *Let $\mathcal{F}$ be a collection of subsets of Ω. $\mathcal{F}$ is said to be an* **algebra** *if*

i) $\emptyset \in \mathcal{F}$.

ii) If $A \in \mathcal{F}$ *then* $A^c \in \mathcal{F}$.

iii) If A and $B \in \mathcal{F}$ *then* $A \cup B \in \mathcal{F}$.

By induction, iii) implies that $\mathcal{F}$ is closed under finite unions, i.e.

$$\text{if } A_1, A_2, \ldots, A_n \in \mathcal{F}, \text{ then } \bigcup_{i=1}^{n} A_i \in \mathcal{F}.$$

Definition 1.5. *If $\mathcal{F}$ is an algebra that is closed under countable unions, i.e.,*

$$\text{if } A_1, A_2, \cdots \in \mathcal{F}, \text{ then } \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}.$$

then $\mathcal{F}$ is said to be a **sigma algebra**. *An element of a sigma algebra $\mathcal{F}$ is said to be an* **event**.

Exercise 1.6. *Show that a sigma algebra is also closed under countable intersections.*

Examples of sigma algebras

Example 1.7. *For any set Ω, the simplest sigma algebra on Ω is the trivial sigma algebra:*

$$\mathcal{F}_0 = \Big\{\emptyset,\ \Omega\Big\}.$$

This sigma algebra is not really useful in practice. It corresponds to an experimental setting where we cannot learn anything about the outcome. The only distinction we can draw is whether something happened or nothing happened.

Example 1.8. *For any set Ω, and any subset $E \subseteq \Omega$,*

$$\mathcal{F}_E = \left\{\emptyset,\ E,\ E^c,\ \Omega\right\}$$

is a sigma algebra on Ω.

This sigma algebra describes an experimental setting in which we can distinguish whether or not the event E has occurred.

Example 1.9. *For any set Ω, e.g. $\Omega = \{1,2,3\}$, the power set of Ω:*

$$\mathcal{F} = \Big\{\emptyset,\ \{1\},\ \{2\},\ \{3\},\ \{1,2\},\ \{1,3\},\ \{2,3\},\ \{1,2,3\}\Big\},$$

the set of all subsets of Ω, is a sigma algebra. When working with a finite or countably infinite state space, this sigma algebra is a natural choice, called the discrete sigma algebra on Ω.

Exercise 1.10. *(An algebra that is not a sigma algebra) Suppose $\Omega = \mathbf{N} = \{1,2,3,\dots\}$ and let $\mathcal{F} = \{A \subseteq \Omega : A \text{ finite or } A^c \text{ finite}\}$.*

i) Identify a subset of Ω that lies in $\mathcal{F}$, and a subset that is not in $\mathcal{F}$.

ii) Show that $\mathcal{F}$ is an algebra.

iii) Show that $\mathcal{F}$ is not a sigma algebra.

The Borel sigma algebra on $\mathbf{R}$

Remark 1.11. *For uncountable state spaces, such as $\mathbf{R}$, we choose to work with a smaller sigma algebra of events than the power set. Practically, this makes sense: whenever we measure a continuous value $x \in \mathbf{R}$, what we really do is observe that the value lies in some interval $(x, x+h)$, where the width h is determined by the precision of the measurement device. This motivates the definition of our usual sigma algebra on $\mathbf{R}$ or other uncountable state spaces, such as $[0,1]$.*

Considering the imprecision of measurement, we at least want the open intervals in $\mathbf{R}$, to be events, i.e. all sets of the form (a,b). In fact, we will work with the smallest sigma algebra that contains all such sets. To be precise about the sense in which our choice is smallest, we need the following result.

Proposition 1.12. *Suppose $\mathcal{F}_i$, $i \in I$ is a non-empty collection of sigma algebras on a set Ω. Then $\cap_{i\in I}\mathcal{F}_i$ is a sigma algebra.*

Proof. *There are three properties to check.*

(1) *Certainly $\emptyset, \Omega \in \mathcal{F}_i$ for each $i \in I$, so these sets are contained in the intersection.*

(2) *If $E \in \cap_{i\in I}\mathcal{F}_i$ then for each $i \in I$, $E \in \mathcal{F}_i$, so $E^c \in \mathcal{F}_i$. Hence $E^c \in \cap_{i\in I}\mathcal{F}_i$.*

(3) *If $E_1, E_2, \ldots \cap_{i\in I} \mathcal{F}_i$, then for each $i \in I$, each set $E_j \in \mathcal{F}_i$, so $\cup_{j=1}^{\infty} E_j \in \cap_{i\in I}\mathcal{F}_i$.*

Hence $\cap_{i\in I}\mathcal{F}_i$ is a sigma algebra.

Definition 1.13. *Let $\mathcal{F}_i$, $i \in I$ be the collection of all sigma algebras that contain all open intervals of $\mathbf{R}$. This collection is clearly non-empty, because the power set of $\mathbf{R}$ is such a sigma algebra. The* **Borel sigma algebra** *$\mathcal{B}$ is defined to be $\cap_{i\in I}\mathcal{F}_i$.*

Remark 1.14. (1) *By construction, $\mathcal{B}$ contains all open intervals along with their complements, countable unions, and countable intersections.*

(2) *By construction, if $\mathcal{F}$ is any sigma algebra containing all intervals of the form above, then $\mathcal{B} \subseteq \mathcal{F}$. In this sense, $\mathcal{B}$ is the smallest sigma algebra containing all intervals.*

(3) *Sets in $\mathcal{B}$ are said to be* **Borel sets**.

(4) *We will not attempt to define probabilities for subsets of $\mathbf{R}$ that are not in $\mathcal{B}$.*

(5) *It is extremely difficult to construct explicitly a set that is not in $\mathcal{B}$.*

Example 1.15. *(Some Borel sets) Using the properties of sigma algebras as needed, for $a, b \in \mathbf{R}$ and $n \in \mathbf{N}$, all sets of the following form are Borel sets.*

- $(a, \infty) = \bigcup_{n=1}^{\infty}(a, n)$,
- $[a, b] = ((-\infty, a) \cup (b, \infty))^c$,
- $a = [a, a]$,
- $\mathbf{N} = \bigcup_{n=1}^{\infty}\{n\}$.

Kolmogorov axioms

Definition 1.16. *(Kolmogorov Axioms) Given a set Ω and a sigma algebra $\mathcal{F}$ on Ω, a* **probability function** *or* **probability measure** *is a function $\Pr : \mathcal{F} \to [0, 1]$ such that*

1. $\Pr(A) \geq 0$, *for all* $A \in \mathcal{F}$.
2. $\Pr(\Omega) = 1$.
3. **Countable additivity**: *If* $A_1, A_2, \ldots \in \mathcal{F}$ *are pairwise disjoint, then*
$$\Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \Pr(A_i).$$

Definition 1.17. *The triple* $(\Omega, \mathcal{F}, \Pr(\cdot))$*, consisting of a sample space* Ω*, a sigma algebra* $\mathcal{F}$ *of subsets of* Ω *and a probability function* $\Pr(\cdot)$ *on* $\mathcal{F}$ *is called a* **probability space**.

All of the standard properties of probability functions follow from these axioms. Suppose, for example, that $A \in \mathcal{F}$, $B \in \mathcal{F}$ and $\{C_1, C_2, \ldots\}$ form a partition of Ω. Recall that this means, $C_i \cap C_j = \emptyset$ for $i \neq j$ and $\cup_{i=1}^{\infty} C_i = \Omega$, with each $C_i \in \mathcal{F}$. Then

1. $\Pr(\emptyset) = 0$.
2. $\Pr(A) \leq 1$.
3. $\Pr(A^c) = 1 - \Pr(A)$.
4. $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$.
5. If $A \subseteq B$ then $\Pr(A) \leq \Pr(B)$.
6. $\Pr(A) = \sum_{i=1}^{\infty} \Pr(A \cap C_i)$.

Countable Additivity and the Continuity Property

Proposition 1.18. *(Continuity property) Suppose* $(\Omega, \mathcal{F}, \Pr)$ *is a probability space. Let* $A_1, A_2, \ldots \in \mathcal{F}$ *be an increasing sequence of events, i.e.* $A_1 \subseteq A_2 \subseteq \ldots$*, so that*
$$A = \bigcup_{n=1}^{\infty} A_n \in \mathcal{F},$$
because $\mathcal{F}$ *is a sigma algebra. Then*
$$\Pr(A) = \lim_{n \to \infty} \Pr(A_n),$$
i.e,
$$\Pr\left(\lim_{n \to \infty} A_n\right) = \lim_{n \to \infty} \Pr(A_n).$$
Similarly, if $B_1, B_2, \ldots \in \mathcal{F}$ *is a decreasing sequence of events, i.e.* $B_1 \supseteq B_2 \supseteq \ldots$*, then*
$$\Pr(B) = \Pr\left(\bigcap_{n=1}^{\infty} B_n\right) = \lim_{n \to \infty} \Pr(B_n).$$

Proof. *Write A as a disjoint union of events:*

$$A = A_1 \cup (A_2 \backslash A_1) \cup (A_3 \backslash A_2) \dots .$$

As we have a disjoint union, we can use the axiom of countable additivity to see that

$$\begin{aligned} \Pr(A) &= \Pr(A_1) + \sum_{i=1}^{\infty} \Pr(A_{i+1} \backslash A_i) \\ &= \Pr(A_1) + \lim_{n\to\infty} \sum_{i=1}^{n-1} \Pr(A_{i+1}) - \Pr(A_i) \\ &= \lim_{n\to\infty} \Pr(A_n), \end{aligned}$$

because of cancellations in the sum. The result for decreasing sequences follows on taking complements.

The result above is in fact equivalent to the assumption of countable additivity, in the following sense.

Proposition 1.19. *Suppose $\mathcal{F}$ is a sigma algebra, and suppose that $\Pr : \mathcal{F} \to [0,1]$ is a finitely additive set function with the property that for any increasing sequence of events, $A_1 \subseteq A_2 \subseteq \dots$*

$$\Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{n\to\infty} \Pr(A_n).$$

Then $\Pr$ is also countably additive.

Proof. *Suppose $A_1, A_2, \dots$ is a sequence of pairwise disjoint sets. Then for each $n \geq 1$,*

$$\Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \Pr\left(\bigcup_{i=1}^{n} A_i \cup \bigcup_{i=n+1}^{\infty} A_i\right)$$

By finite additivity, we see that

$$\Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{n} \Pr\left(A_i\right) + \Pr\left(\bigcup_{i=n+1}^{\infty} A_i\right).$$

Now, define $B_n = \bigcup_{i=n+1}^{\infty} A_i$, then we have a decreasing sequence of events $B_1 \supseteq B_2 \supseteq \dots$,
and, since the sets A_i are disjoint, we have

$$\bigcap_{n=1}^{\infty} B_n = \emptyset,$$

because any element of this intersection would be contained in infinitely many of the A_i. Taking the limit $n \to \infty$, and using the result for decreasing sequences, we see that

$$\Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \Pr(A_i) + \lim_{n\to\infty} \Pr(B_n) = \sum_{i=1}^{\infty} \Pr(A_i) + \Pr(\emptyset) = \sum_{i=1}^{\infty} \Pr(A_i).$$

Lebesgue measure

We have a natural sense of what it means to define **uniform** probability on an interval such as $[0,1]$. The corresponding probability function assigns to any interval $(a,b) \subseteq [0,1]$ the probability $b-a$, proportional to its length. This is the probability measure that corresponds to the idea of sampling at a number *at random* from $[0,1]$.

It takes quite a lot of effort to construct this probability function, which is known as **Lebesgue measure** . It is defined on a sigma algebra, the **Lebesgue sigma algebra**, which contains the Borel sigma algebra $\mathcal{B}$ as a subset. It is difficult to construct a set that is not contained within the Lebesgue sigma algebra, and no 'naturally occurring' sets ever fail to be Lebesgue measurable.

In a sense, Lebesgue measure is the only probability function we need, because any other probability functions of interest can be obtained from Lebesgue measure by a transformation. See Section 4.6 of Proschan and Shaw for details.

The Vitali set (non-examinable)

As an illustration, we will give an example of the construction of a set that is *not* Lebesgue measurable (and so is not in $\mathcal{B}$ either). This construction requires the axiom of choice. Let S^1 be the unit circle, parameterized by angle as $[0, 2\pi)$, and suppose we use Lebesgue measure to define a uniform probability measure on S^1, i.e. the probability of an interval $\Pr((a,b))$ is given by $(b-a)/2\pi$. Note that this measure is rotation-invariant, in the sense that $(a+x, b+x)$, thought of as a subset of S^1, is a rotation of (a,b) through the angle x.

Define an equivalence relation on S^1 by $x \sim y$ if and only if $x-y$ is a rational multiple of π. Let A be a **transversal** for this equivalence relation, i.e. a set containing exactly one representative of each equivalence class. (This is the step that requires the axiom of choice.)

Suppose now that $x_1, x_2, \ldots$ is an enumeration of the rational angles in $[0, 2\pi)$. Then the sets $A_i = \{a + x_i : a \in A\}$ can be seen to be

pairwise disjoint and such that

$$S^1 = \bigcup_{i=1}^{\infty} A_i.$$

Note that A_i is just a rotation of the set A through the rational angle x_i. So, if we could define $\Pr(A)$ using Lebesgue measure, as above for intervals, then $\Pr(A) = \Pr(A_i)$. But then by countable additivity of Pr,

$$1 = \Pr(S^1) = \Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \Pr(A_i) = \sum_{i=1}^{\infty} \Pr(A).$$

There is no value that can be assigned to $\Pr(A)$ to make this equality hold, and so we have reached a contradiction. It cannot be the case that A is a set in the Lebesgue sigma algebra.

The interpretation is that, despite our strong intuition, the idea of 'uniform probability' embodied by Lebesgue measure does not extend to arbitrary subsets of the real line.

Elementary probability results (Review)

Definition 1.20. *If A and $B \in \mathcal{F}$ with $\Pr(B) > 0$, the* **conditional probability** *of A given B is*

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

Definition 1.21. *Two events are* **independent** *if*

$$\Pr(A \cap B) = \Pr(A)\Pr(B).$$

If $\Pr(B) > 0$, A and B independent means

$$\Pr(A|B) = \Pr(A).$$

Definition 1.22. *A collection of events $A_1, \ldots, A_n \in \mathcal{F}$ is* **mutually independent** *if for any subcollection $A_{i_1}, \ldots, A_{i_k}$, we have*

$$\Pr\left(\bigcap_{j=1}^{k} A_{i_j}\right) = \prod_{j=1}^{k} \Pr(A_{i_j}).$$

Pairwise independence is not enough, nor is a factorization for the entire collection.

Example 1.23. *Pairwise independence without mutual independence. Consider an experiment in which a coin is flipped twice, and the outcome*

recorded. Define the following events:

$$A = \{HH,\ HT\},$$
$$B = \{HH,\ TH\},$$
$$C = \{HT,\ TH\}.$$

Clearly, we have

$$\Pr(A) = \Pr(B) = \Pr(C) = \frac{1}{2},$$

and

$$\Pr(A \cap B) = \Pr(\{HH\}) = \frac{1}{4} = \Pr(A)\Pr(B)$$
$$\Pr(A \cap C) = \Pr(\{HT\}) = \frac{1}{4} = \Pr(A)\Pr(C)$$
$$\Pr(B \cap C) = \Pr(\{TH\}) = \frac{1}{4} = \Pr(B)\Pr(C),$$

but these events are not independent, because

$$\Pr(A \cap B \cap C) = \Pr(\emptyset) = 0 \neq \Pr(A)\Pr(B)\Pr(C).$$

Example 1.24. *Factorization for the entire collection, without pairwise independence.*

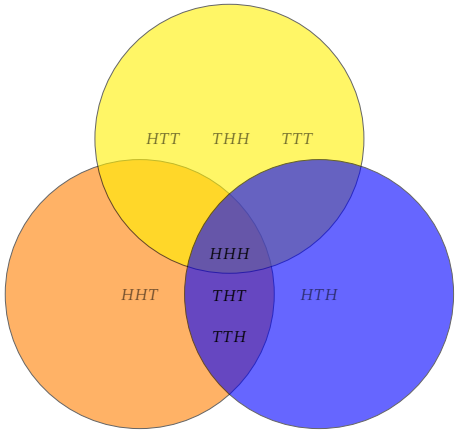


Figure 1.2: The three events shown are *not* pairwise independent, and so not independent.

Consider an experiment in which a coin is flipped three times, and the outcome recorded. Define the following events, which exhaust the sample space, depicted in the Figure 1.2:

$$A = \{HHH,\ HHT,\ THT,\ TTH\},$$
$$B = \{HHH,\ HTH,\ THT,\ TTH\},$$
$$C = \{HHH,\ HTT,\ THH,\ TTT\}.$$

Clearly, we have

$$\Pr(A) = \Pr(B) = \Pr(C) = \frac{1}{2},$$

and

$$\Pr(A \cap B \cap C) = \Pr(\{HHH\}) = \frac{1}{8} = \Pr(A)\Pr(B)\Pr(C),$$

but these events are not independent, because

$$\Pr(A \cap B) = \Pr(\{HHH,\ THT,\ TTH\}) = \frac{3}{8} \neq \Pr(A)\Pr(B).$$

2

Random Variables

Introduction

Suppose we have a probability space $(\Omega, \mathcal{F}, Pr)$. A random variable on this space is a function $X : \Omega \to \mathbf{R}$. For practical purposes, it is a way of associating an outcome with a real number. So far, we have used the sample space as an abstraction of the idea of possible outcomes of an experiment. Another good mental picture to keep in mind, particularly for statistics, is to think of Ω as a population of individuals, with many different attributes. A random variable is a measurement of a particular attribute, for each individual in the population.

We have seen that a natural collection of events on $\mathbf{R}$ is the Borel sigma algebra $\mathcal{B}$, the smallest sigma algebra that contains all open intervals. This is because measurements are always of finite precision, and so the measurement of a real value is actually an observation that the value lies within some interval.

The probability function Pr has domain $\mathcal{F}$, so if we want to make probability statements about X, e.g. $\Pr(X \in B)$ for some $B \in \mathcal{B}$, then we need to ensure that the **pre-image** $X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\}$, is an event in $\mathcal{F}$. This motivates our formal definition.

Definition 2.1. *A* **random variable** *on the probability space* $(\Omega, \mathcal{F}, \Pr)$ *is a function,*

$$X : \Omega \to \mathbf{R},$$

such that for every Borel set $B \in \mathcal{B}$, $X^{-1}(B) \in \mathcal{F}$. *This condition can be summarized by saying that a random variable is an* $\mathcal{F}$*- measurable function.*

We can also define random vectors , $X : \Omega \to \mathbf{R}^n$, *and complex random variables,* $X : \Omega \to \mathbf{C}$, *in an analogous way.*

Example 2.2. *Suppose we flip a coin twice, so that* $\Omega = \{HH, HT, TH, TT\}$. *Let the function* $X : \Omega \to \mathbf{R}$ *count the number of heads observed, i.e.*

$$\begin{aligned} X(HH) &= 2 \\ X(HT) = X(TH) &= 1 \\ X(TT) &= 0. \end{aligned}$$

For discrete problems, we typically take $\mathcal{F}$ *to be the power set of* Ω*, the set of all subsets of* Ω*. In this case,* X *is certainly a random variable, because for any* $B \in \mathcal{B}$, $X^{-1}(B) \subseteq \Omega$, *and* $\mathcal{F}$ *contains all subsets of* Ω.

However, suppose instead we took as our sigma algebra

$$\mathcal{F}_1 = \{\emptyset, \{HH, TT\}, \{HT, TH\}, \Omega\}.$$

This would correspond to an experiment where we can only distinguish whether the two flips were the same or different. For example, suppose that instead of flipping coins with labelled faces, we flip unlabelled magnetic discs, with H and T corresponding to North and South poles. We can easily determine whether the discs have landed with the same pole facing up, by seeing whether the discs mutually attract or repel. We could not determine the specific polarities without more work. With respect to $\mathcal{F}_1$, X *is not a random variable. To see this explicitly, note that* $\{2\} \in \mathcal{B}$, *but* $X^{-1}(\{2\}) = \{HH\} \notin \mathcal{F}_1$.

A random variable X induces a probability function $\Pr_X : \mathcal{B} \to [0, 1]$, as follows.

Definition 2.3. *For any Borel set* $B \in \mathcal{B}$, *let*

$$\Pr_X(B) = \Pr(X^{-1}(B)) = \Pr\left(\{\omega \in \Omega : X(\omega) \in B\}\right).$$

$\Pr_X$ *is called the* **distribution** *of* X. *Abusing notation slightly, we usually write* $\Pr_X(B)$ *as* $\Pr(X \in B)$.

Example 2.4. *Let* $(\Omega, \mathcal{F}, \Pr)$ *be a probability space, then define* $X : \Omega \to \mathbb{R}$ *to be the constant function* $X(\omega) = c$ *for some* $c \in \mathbb{R}$. *To show that* X *is a random variable, we have to show that* $X^{-1}(B) \in \mathcal{F}$ *, for any Borel set* $B \in \mathcal{B}$. *There are clearly two cases to consider, according to whether or not* $c \in B$:

$$X^{-1}(B) = \begin{cases} \Omega & c \in B \\ \emptyset & c \notin B. \end{cases}$$

As $\emptyset \in \mathcal{F}$ *and* $\Omega \in \mathcal{F}$, *we see that* X *is measurable. The distribution of* X *is given by*

$$\Pr(X \in B) = \begin{cases} 1 & c \in B \\ 0 & c \notin B. \end{cases}$$

Example 2.5. *Let $(\Omega, \mathcal{F}, \Pr)$ be a probability space, then for any event $A \in \mathcal{F}$, the indicator random variable $\mathbf{1}_A : \Omega \to \mathbb{R}$ is defined*

$$\mathbf{1}_A(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A. \end{cases}$$

To show that $\mathbf{1}_A$ is a random variable, we have to show that $\mathbf{1}_A^{-1}(B) \in \mathcal{F}$, for any Borel set $B \in \mathcal{B}$. There are several cases to consider:

$$\mathbf{1}_A^{-1}(B) = \begin{cases} \emptyset & 0 \notin B, 1 \notin B \\ A^c & 0 \in B, 1 \notin B \\ A & 0 \notin B, 1 \in B \\ \Omega & 0 \in B, 1 \in B. \end{cases}$$

In each case $\mathbf{1}_A^{-1}(B) \in \mathcal{F}$, so $\mathbf{1}_A$ defines a random variable. If $\Pr(A) = p$, then the distribution of X is

$$\Pr(\mathbf{1}_A \in B) = \begin{cases} 0 & 0 \notin B, 1 \notin B \\ 1-p & 0 \in B, 1 \notin B \\ p & 0 \notin B, 1 \in B \\ 1 & 0 \in B, 1 \in B. \end{cases}$$

Proposition 2.6. *We need some guarantees that we can readily make new random variables from existing ones. We have the following technical result, which we shall always assume.*

1. *If X is a random variable, then so are $X + a$, aX, X^2, where $a \in \mathbf{R}$.*
2. *If X and Y are random variables, then so are $X + Y$ and XY.*
3. *If $X_1, X_2, \ldots$ is a sequence of random variables such that for all $\omega \in \Omega$, $X(\omega) = lim_{n\to\infty} X_n(\omega)$ exists, then X is a random variable.*

Proof. *Not examinable. See propositions 4.7 and 4.9 in Proschan and Shaw.*

Definition 2.7. *We say X and Y are* **identically distributed** *if $\Pr(X \in B) = \Pr(Y \in B)$ for all $B \in \mathcal{B}$.*

Example 2.8. *Suppose a fair coin is flipped independently five times, so that Ω is the collection of sequences of length 5 from the alphabet $\{H, T\}$, and let $\mathcal{F}$ be the discrete sigma algebra on Ω.*

Let X be number of heads and Y be number of tails. By symmetry, X and Y are identically distributed, even though $X(\omega) \neq Y(\omega)$ for all $\omega \in \Omega$.

In general, checking conditions for all Borel sets is cumbersome. We define a more useful condition for measurability.

Proposition 2.9. *If $(\Omega, \mathcal{F}, \Pr)$ is a probability space, then $X : \Omega \to \mathbf{R}$ is a random variable if and only if for all $x \in \mathbf{R}$,*

$$\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}.$$

Proof. For X to be measurable, it is clearly necessary that $\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}$, because $(-\infty, x]$ is a Borel set.

To check that the condition is also sufficient, we suppose that X is a function such that $X^{-1}(-\infty, x] \in \mathcal{F}$ for all $x \in \mathbf{R}$. The collection $\mathcal{A} = \{B \in \mathcal{B} : X^{-1}(B) \in \mathcal{F}\}$ can easily be seen to be a sigma algebra (exercise). Hence, it is enough to show that $(a, b) \in \mathcal{A}$ for all $a, b \in \mathbf{R}$. Then we must have $\mathcal{A} = \mathcal{B}$, because $\mathcal{B}$ is the smallest sigma algebra to contain all open intervals.
To see this, note that for $a < b$, we have

$$(a, b) = (-\infty, b) \cap (a, \infty)$$

so it is enough to show that $(-\infty, b), (a, \infty) \in \mathcal{A}$.

To do this, note that $(a, \infty) = (-\infty, a]^c \in \mathcal{A}$. Then, write

$$(-\infty, b) = \bigcup_{n=1}^{\infty} \left(-\infty, b - \frac{1}{n}\right] \in \mathcal{A}.$$

Hence $(a, b) \in \mathcal{A}$, hence $\mathcal{A} = \mathcal{B}$.

Definition 2.10. *The* **cumulative distribution function** *(CDF) of a random variable X is a function $F_X : \mathbf{R} \to [0, 1]$ defined by*

$$F_X(x) = \Pr(X \leq x).$$

Proposition 2.11. *X and Y are identically distributed if and only if $F_X(x) = F_Y(x)$ for all $x \in \mathbf{R}$.*

Proof. *Not examinable. See proposition 4.22 in Proschan and Shaw.*

Definition 2.12. *We write $x_n \downarrow x$ if (x_n) is a sequence (weakly) monotonically decreasing to the limit x, and $x_n \uparrow x$ if (x_n) is (weakly) monotonically increasing to x.*

Proposition 2.13. *If X is a random variable with CDF F_X, then*

1. *$F_X(x)$ is non-decreasing.*
2. $\lim_{x\to-\infty} F_X(x) = 0$; $\lim_{x\to+\infty} F_X(x) = 1$.
3. $\lim_{x\downarrow x_0} F_X(x) = F_X(x_0)$. *This says that F is continuous from the right.*

Proof.

1. If $x \leq y$ then $(-\infty, x] \subseteq (-\infty, y]$ so $\Pr(X \leq x) \leq \Pr(X \leq y)$.
2. For any sequence $(x_n)_{n\geq 1}$ such that $x_n \uparrow \infty$, define the increasing sequence of events $A_n = \{\omega \in \Omega : X(\omega) \in (-\infty, x_n]\}$. By the continuity property,

$$\lim_{n\to\infty} \Pr(X \leq x_n) = \lim_{n\to\infty} \Pr(A_n) = \Pr\left(\bigcup_{n=1}^{\infty} A_n\right) = \Pr(X \in \mathbf{R}) = 1.$$

   A similar argument with a decreasing sequence shows that $\lim_{x\to-\infty} F_X(x) = 0$.
3. Let $(x_n)_{n\geq 1}$ be a sequence such that $x_n \downarrow x$ as $n \to \infty$. Define $B_n = \{\omega \in \Omega : X(\omega) \leq x_n\}$. Then B_n is a decreasing sequence of events so by the continuity property,

$$\Pr(B_n) \downarrow \Pr(B),$$

   where

$$B = \bigcap_{n=1}^{\infty} B_n = \{\omega \in \Omega : -\infty < X(\omega) \leq x\}.$$

Types of Random Variable

Perhaps the simplest, but most trivial, type of random is the **constant random variable**. For $a \in \mathbf{R}$, define the **point mass** CDF

$$\delta_a(x) = \begin{cases} 0 & x < a \\ 1 & x \geq a. \end{cases}$$

Then δ_a is the CDF of a random variable X such that $\Pr(X = a) = 1$.

Definition 2.14. *Let X be a random variable with CDF F_X.*

(1) If there exist sequences of real values $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$ where $b_i > 0$ and $\sum_{i=1}^{\infty} b_i = 1$, and F_X is such that

$$F_X(x) = \sum_{i=1}^{\infty} b_i \delta_{a_i}(x),$$

then X is a **discrete** *random variable.. The* **probability mass function** *(PMF) of a discrete random variable X is $f_X(x) = \Pr(X = x)$.*

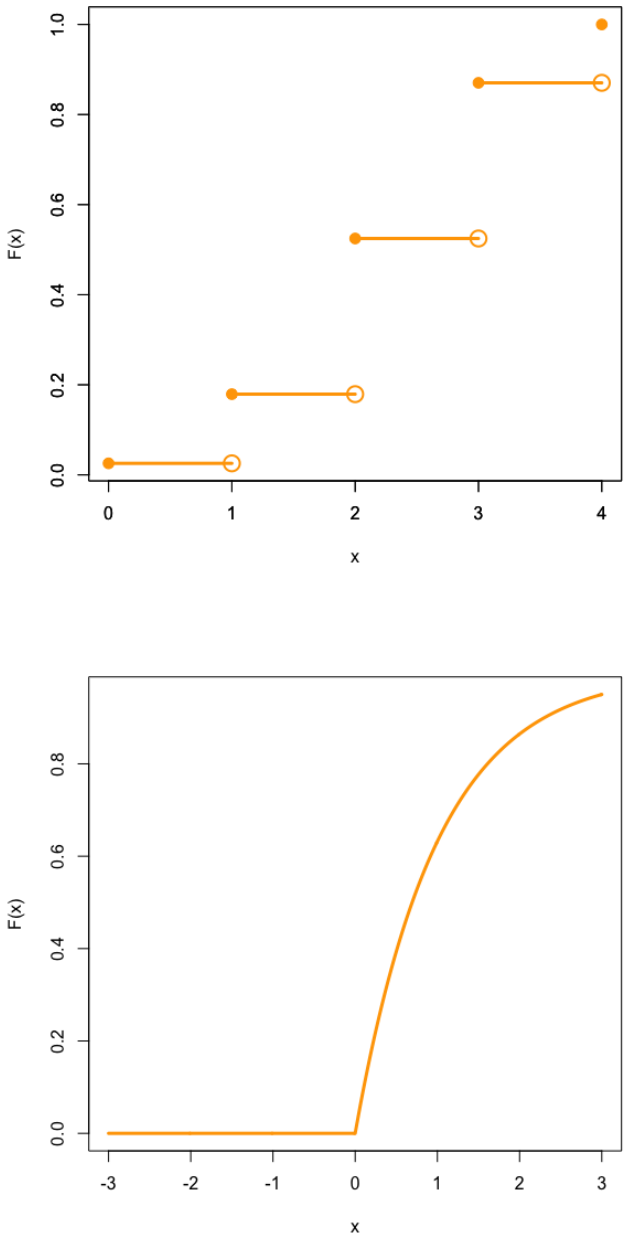


Figure 2.1: The top panel shows the CDF of a $\text{Bin}(4, 0.6)$ variable in the range $[0, 4]$. Note the jump discontinuities at each integer. The bottom panel shows the CDF of an $\text{Exp}(1)$ variable in the range $[-3, 3]$. Note that the function is continuous.

(2) *If F_X is a continuous function, then X is a* **continuous** *random variable.*

(3) *If X is a continuous random variable such that there exists a function $f_X : \mathbb{R} \to \mathbb{R}$ satisfying*

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t$$

for all $x \in \mathbf{R}$, then X is an **absolutely continuous** *random variable. Such a function f_X is said to be a* **probability density function** *for X.*

In general, F_X may be neither continuous nor discrete. Practical examples are easy to find in statistics, e.g. let X be the volume of beer consumed over 24 hours by a UK adult. For some proportion of the population, the volume is precisely zero. But conditional on having consumed a positive amount, the variable is continuous. So that $\Pr(X = x) = 0$ for all $x > 0$.

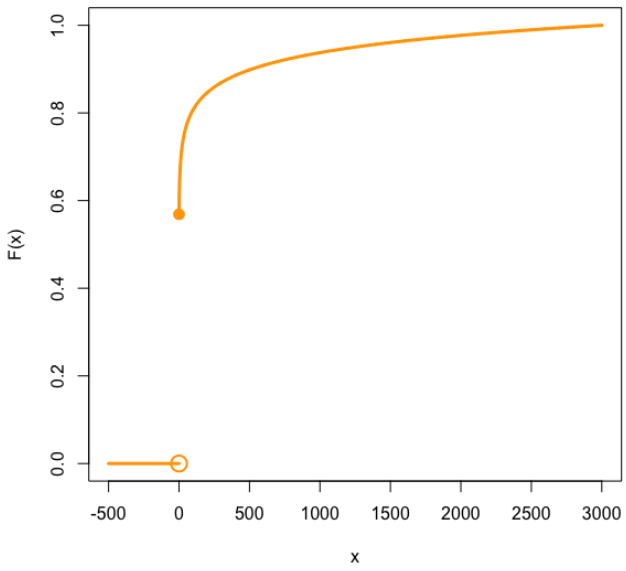


Figure 2.2: The CDF of the random variable X, which shows a simple model for the volume of beer consumed over 24 hours by a randomly selected UK adult. Note that the CDF is in principle defined for all $x \in \mathbf{R}$, although only non-negative values of the variable make sense. Note the discontinuity at 0: a positive proportion of subjects consume no beer.

Remark 2.15. *In a probability theory course, more precise statements can be made about the character of distributions. In particular, the* **Lebesgue decomposition theorem** *states that any cumulative distribution function F can be written uniquely as*

$$F(x) = \alpha F_c(x) + \beta F_d(x) + \gamma F_s(x),$$

for $\alpha, \beta, \gamma \geq 0$ and $\alpha + \beta + \gamma = 1$, in which F_d is a discrete distribution function, F_c is an absolutely continuous distribution function, and F_s is the CDF of a singular continuous distribution.

A **singular continuous** *distribution has a continuous CDF, but there is no corresponding probability density function. An example of such a singular distribution is the Cantor distribution, discussed below. Singular distributions will not feature in this course, beyond this non-examinable example. If you continue to think of a continuous random variable as having a probability density function, you will encounter no problems in this course.*

Example 2.16. *If F is the CDF in Figure 2.2, we can write*

$$F(x) = \alpha\delta_0(x) + (1-\alpha)F_c(x),$$

where $\delta_0(x)$ is a point mass at zero, α is the proportion of individuals who have consumed no beer, and F_c is the absolutely continuous CDF of the amount consumed by those who have consumed a non-zero amount.

Proposition 2.17. *For any random variable, $\Pr(X < x) = \lim_{x_n \uparrow x} \Pr(X \leq x_n)$, for any strictly increasing sequence $x_n \uparrow x$.*

Proof. *Define the increasing sequence of events*

$$A_n = \{\omega \in \Omega : -\infty < X(\omega) \leq x_n\}.$$

Then for any $y < x$, since $x_n \uparrow x$ there exists $n \in \mathbf{N}$ such that $y \leq x_n$, so $y \in A_n$. Hence,

$$\bigcup_{n=1}^{\infty} A_n = \{\omega \in \Omega : -\infty < X(\omega) < x\},$$

so the result follows.

Proposition 2.18. *If X is a continuous random variable, i.e. if F_X is a continuous function, then for all $x \in \mathbf{R}$, $\Pr(X = x) = 0$.*

Proof. *Applying the previous result, using, say, $x_n = x - \frac{1}{n}$,*

$$\begin{aligned}\Pr(X = x) &= \Pr(X \leq x) - \Pr(X < x) = \Pr(X \leq x) - \lim_{n\to\infty} \Pr\left(X \leq x - \frac{1}{n}\right) \\ &= F_X(x) - \lim_{n\to\infty} F_X(x - \frac{1}{n}) = 0,\end{aligned}$$

by the continuity of F_X.

Remark 2.19. *The probability density function of an absolutely continuous random variable X is* **any** *function g that satisfies*

$$F_X(x) = \int_{-\infty}^{x} g(t)\,\mathrm{d}t, \quad \text{for all } x.$$

If the function f_X defined by

$$\frac{\mathrm{d}}{\mathrm{d}x}F_X(x) = f_X(x)$$

is continuous, then by the fundamental theorem of calculus, f_X is **one** *function that satisfies the definition of a PDF. But we could alter $f_X(x)$ at any single point without changing $F_X(x)$. So PDFs are not unique.*

The Cantor distribution (non-examinable)

We now construct an example of a **singular continuous** distribution, the **Cantor distribution**. It will have a continuous CDF, but there will be no corresponding PDF.

We begin with an iterative construction, which is illustrated in Figure 2.5. Start with $[0,1]$, and remove the open middle third $\left(\frac{1}{3}, \frac{2}{3}\right)$. Then remove the open middle third of the two intervals that remain. We can imagine repeating this process indefinitely. But note that however far we go, some elements of $[0,1]$ would never be removed. These form the **Cantor set**.

More precisely, define $C_0 = [0,1]$ and for $n \geq 1$, set

$$C_n = \frac{C_{n-1}}{3} \cup \left(\frac{2}{3} + \frac{C_{n-1}}{3}\right).$$

After step n, we have removed $1 + 2 + \ldots + 2^{n-1} = 2^n - 1$ disjoint open intervals. This then gives a decreasing sequence of sets $C_0 \supseteq C_1 \supseteq \ldots$, as shown in Figure 2.5. We see that C_n is just the union of 2^n disjoint closed intervals, each of length 3^{-n}.

The Cantor set is defined to be $C = \bigcap_{n=1}^{\infty} C_n$. Note that $C \in \mathcal{B}$, as a countable intersection of Borel sets. It therefore makes sense to ask for the probability that random variables are in C.

So suppose $U \sim \mathrm{Unif}[0,1]$. What is $\Pr(U \in C)$? To determine this, apply the continuity property to the decreasing sequence (C_n).

$$\Pr(C) = \Pr(\lim_{n\to\infty} C_n) = \lim_{n\to\infty} \Pr(C_n) = \lim_{n\to\infty} \frac{2^n}{3^n} = 0.$$

This says that C is a **null set**: given any $\epsilon > 0$, there are intervals $I_1, I_2, \ldots$ of $[0,1]$ whose total length is at most ϵ, and for which $C \subseteq \bigcup_n I_n$. Nonetheless, C still contains uncountably many real numbers!

One explicit way of working with the Cantor set is to represent each $x \in [0,1]$ as a **ternary expansion**. This is the same principle as a decimal or binary expansion, but using base 3. Explicitly, write

$$x = \sum_{n=1}^{\infty} \frac{a_n}{3^n}, \qquad a_n \in \{0, 1, 2\}.$$

Numbers with $a_1 = 1$ lie in the middle third of the interval - all of these are removed in the first stage. All numbers with $a_2 = 1$ are removed in the next stage, and in general all numbers with $a_k = 1$ are removed in stage k. So the Cantor set is the collection of all real numbers whose ternary expansions contain only 0 and 2. From this representation, we can now use essentially the same diagonal argument as used by Cantor for all of $\mathbf{R}$ to see that C contains uncountably many elements.

We now define the Cantor distribution by specifying a distribution function F on $[0, 1]$. We first specify F for $x \in C$ by

$$F(x) = \sum_{n=1}^{\infty} \frac{a_n}{2^{n+1}}.$$

Then we see easily that

$$F(0) = 0, \qquad F(1) = 1,$$

and moreover, $F(x) \leq F(y)$ whenever $x \leq y$. To extend F to all of $[0, 1]$, take $x \in C\,[0, 1]$ and define

$$F(x) = \sup\{F(y) : y \in C, y < x\}.$$

Finally, we extend the domain of F to all of $\mathbf{R}$ by $F(x) = 0$ for $x < 0$ and $F(x) = 1$ for $x > 1$. This CDF is shown in the lower panel of Figure 2.5. It can be shown to be a continuous function.

Note that by construction F is constant on each of the intervals that are removed. But these intervals cover almost all of $[0, 1]$, in the following sense. At stage n of the construction of C, we removed 2^{n-1} intervals, each of length $\frac{1}{3^n}$. This says that the total removed length by stage n is $\frac{1}{3}\sum_{k=1}^{n} \left(\frac{2}{3}\right)^{k-1} = 1 - \left(\frac{2}{3}\right)^n$, consistent with the computation for C_n above. So the total fraction of $[0, 1]$ that is removed is

$$\frac{1}{3}\sum_{k=1}^{\infty} \left(\frac{2}{3}\right)^{k-1} = \frac{1}{3}\frac{1}{1 - \frac{2}{3}} = 1.$$

So we see that the derivative $F'(x)$ must be zero on almost all of $[0, 1]$. If there were a corresponding probability density function f, we would have $f(x) = 0$ for $x \notin C$, so that

$$\int_{-\infty}^{\infty} f(x)\,dx = 0,$$

in contradiction to the fact that $F(x) = 1$ for $x \geq 1$.

Transformations of random variables

Suppose we have a discrete random variable X whose distribution is given by

x	-1	0	1
$\Pr(X = x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

In this simple example, we can easily determine the distribution of the transformed random variable $Y = X^2$ directly from the probability mass function.

Clearly the only values that Y assumes with positive probability are 0 and 1. We determine these probabilities by direct calculation

$$\Pr(Y = 0) = \Pr(X^2 = 0) = \Pr(X = 0) = \frac{1}{2},$$

For $Y = 1$, note that there are two corresponding possibilities for X.

$$\Pr(Y = 1) = \Pr(X^2 = 1) = \Pr(X \in \{-1, 1\}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.$$

For all but the very simplest problems, it is more straightforward to determine the distribution of a function of a random variable by determining its CDF. An example illustrates this point.

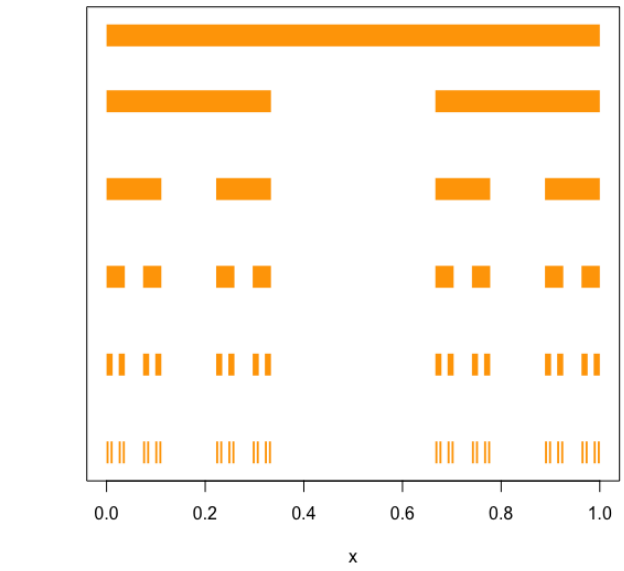
Example 2.20. *Suppose $X \sim \text{Unif}[0, 2\pi]$, so that X has probability mass function*

$$f_X(x) = \begin{cases} \frac{1}{2\pi}, & x \in [0, 2\pi] \\ 0 & \textit{otherwise,} \end{cases}$$

and cumulative distribution function

$$F_X(x) = \begin{cases} 0, & x < 0 \\ \frac{x}{2\pi}, & x \in [0, 2\pi] \\ 1 & x > 2\pi. \end{cases}$$

Consider the random variable $Y = \sin X$. Clearly Y takes values in $[-1, 1]$, but the transformation is not one-to-one, as can be seen in Figure 2.4. Then



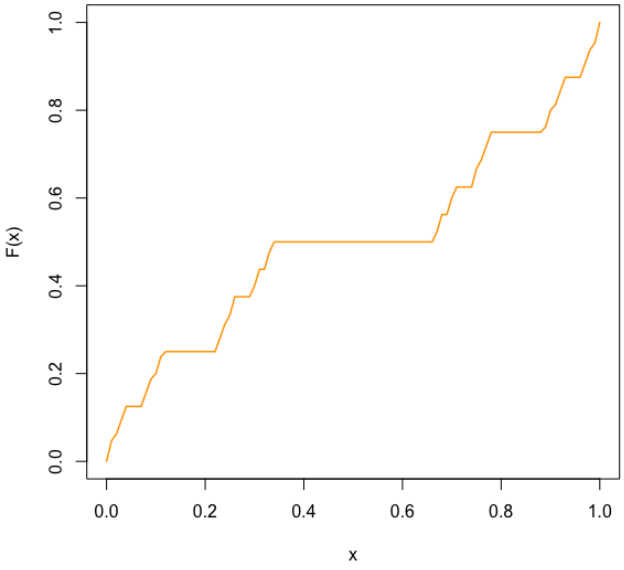


Figure 2.3: Consider an iterative process acting on the interval $[0,1]$. At each stage, delete the middle third of each interval that remains. However many times this procedure is repeated, there are some points that will never be deleted. These form the Cantor set. The CDF of the Cantor distribution is shown below. While the CDF is a continuous function, it is nowhere differentiable, and so does not have a probability density function.

$$\begin{aligned}\Pr(Y \le y) &= \Pr(\sin X \le y) = \Pr(X \le \sin^{-1}(y)) + \Pr(X \ge \pi - \sin^{-1}(y)) \\ &= \frac{1}{2\pi}\sin^{-1}(y) + 1 - \frac{1}{2\pi}\left(\pi - \sin^{-1}(y)\right) \\ &= \frac{1}{2} + \frac{1}{\pi}\sin^{-1}(y).\end{aligned}$$

The distribution of Y is now determined. We can specify its density function by differentiating F_Y:

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{1}{\pi\sqrt{1-y^2}} \qquad -1 \le y \le 1.$$

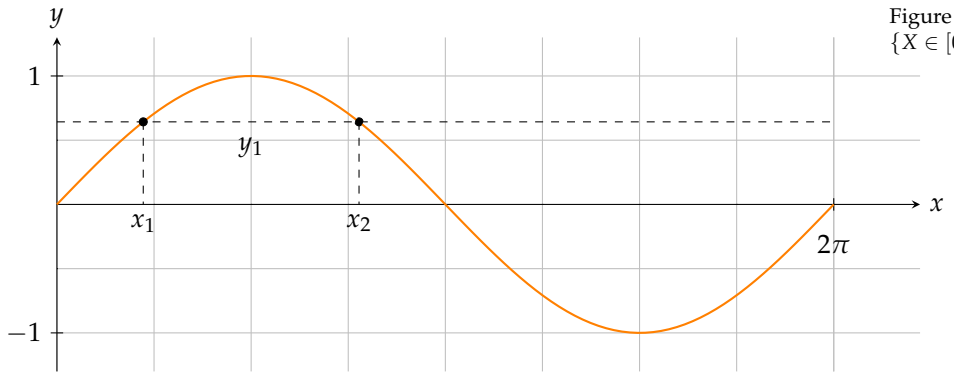


Figure 2.4: If $Y = \sin X$, $\{Y \le y_1\} = \{X \in [0, x_1]\} \cup \{X \in [x_2, 2\pi]\}$.

One-to-one transformations

Proposition 2.21. *Suppose X is an absolutely continuous random variable with probability density function f_X, and $g : \mathbf{R} \to \mathbf{R}$ is a strictly monotonic, differentiable function. Then $Y = g(X)$ has probability density function*

$$f_Y(y) = f_X(g^{-1}(y))\left|\frac{d\,g^{-1}(y)}{dy}\right| \qquad y \in \mathbf{R}.$$

Proof. *Suppose first that g is monotonic increasing. Note that $g(X) \le y$ if and only if $X \le g^{-1}(y)$. Then*

$$\Pr(Y \le y) = \Pr(g(X) \le y) = \Pr(X \le g^{-1}(y)) = F_X(g^{-1}(y)).$$

Then we get f_Y from differentiating F_Y:

$$f_Y(y) = \frac{d}{dy}F_X(g^{-1}(y)) = f_X(g^{-1}(y))\frac{d\,g^{-1}(y)}{dy},$$

by the chain rule.

The argument for decreasing g is similar. Note that $g(X) \leq y$ if and only if $X \geq g^{-1}(y)$. Then

$$\Pr(Y \leq y) = \Pr(g(X) \leq y) = \Pr(X \geq g^{-1}(y)) = 1 - F_X(g^{-1}(y)),$$

so that on differentiating,

$$f_Y(y) = \frac{d}{dy}\left(1 - F_X(g^{-1}(y))\right) = -f_X(g^{-1}(y))\frac{d\,g^{-1}(y)}{dy} = \left|f_X(g^{-1}(y))\frac{d\,g^{-1}(y)}{dy}\right|,$$

noting that, as a decreasing function, g^{-1} has negative derivative.

Remark 2.22. *One way to understand this transformation formula is to consider a simple linear transformation $Y = aX + b$.*

We can consider the probability density function $f_X(x)$ as giving the rate of increase in the probability that X lies in a small interval around x with the width h of the interval:

$$f_X(x)h = \Pr(X \in (x, x+h)) + o(h),$$

where the term $o(h) \to$ is best thought of as an error much smaller than h as $h \to 0$.

If we make the transformation $Y = aX + b$, for $a > 0$, then the probability density for Y satisfies

$$\begin{aligned} f_Y(y)h &= \Pr(Y \in (y, y+h)) + o(h) = \Pr(aX + b \in (y, y+h)) + o(h) \\ &= \Pr\left(X \in \left(\frac{y-b}{a}, \frac{y+h-b}{a}\right)\right) + o(h). \end{aligned}$$

Writing this in terms of f_X gives

$$f_Y(y)h = f_X\left(\frac{y-b}{a}\right)\frac{h}{a} + o(h)$$

since the interval is of length $\frac{h}{a}$. Equating terms of the same order in h then gives the density for Y:

$$f_Y(y) = f_X\left(\frac{y-b}{a}\right)\frac{1}{a}$$

Proposition 2.23. *Suppose $(\Omega, \mathcal{F}, \Pr)$ is a probability space, X is a random variable and $g : \mathbf{R} \to \mathbf{R}$ is $\mathcal{B}$-measurable, i.e. $g^{-1}(B) \in \mathcal{B}$ for all $B \in \mathcal{B}$. Then $Y = g(X)$ is also a random variable.*

Proof. *We need to show that $Y^{-1}(B) \in \mathcal{F}$ for all $B \in \mathcal{B}$. Now,*

$$Y^{-1}(B) = \{\omega \in \Omega : g(X(\omega)) \in B\} = \{\omega \in \Omega : X(\omega) \in g^{-1}(B)\}.$$

But, since g is $\mathcal{B}$-measurable, $g^{-1}(B) \in \mathcal{B}$, so that, since X is $\mathcal{F}$-measurable,

$$Y^{-1}(B) = X^{-1}(g^{-1}(B)) \in \mathcal{F}.$$

Scale and Location families

Suppose Z is a random variable with probability density function f_Z. We can manufacture an entire family of random variables using Z, e.g.

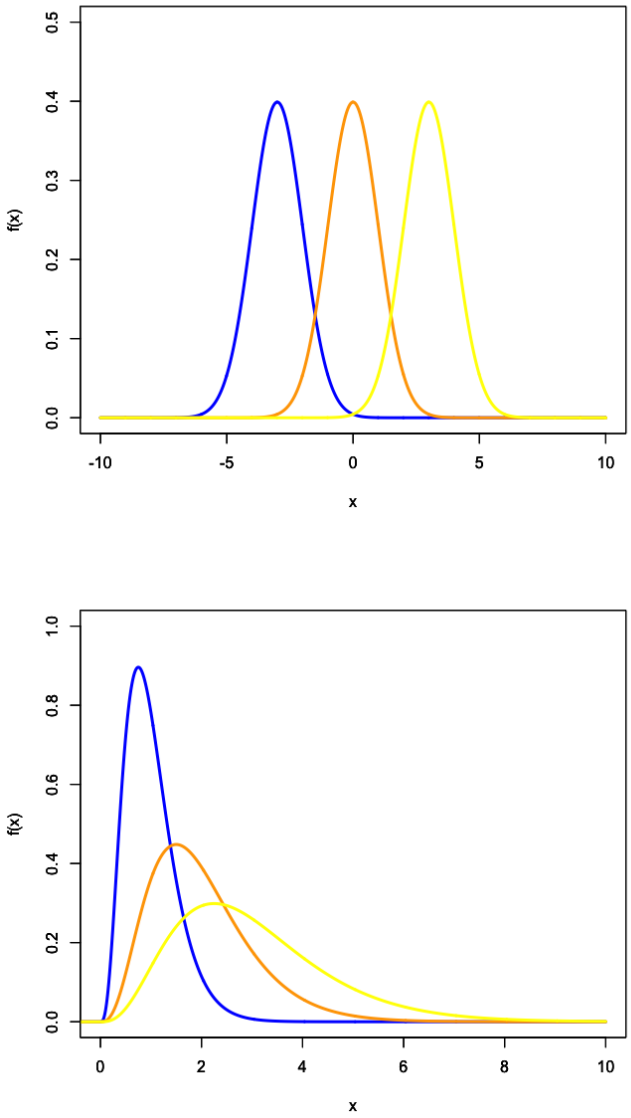


Figure 2.5: Three members of a normal location family (top) and a gamma scale family (bottom)

Location family Define $X = \mu + Z$, which has PDF

$$f(x|\mu) = f_Z(x - \mu).$$

This corresponds to a shift of location by μ.

Example 2.24. *Let $Z \sim N(0,1)$. Then*

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right), \qquad z \in \mathbf{R}.$$

Then the random variable $X = \mu + Z \sim N(\mu, 1)$. Its density is

$$f_X(x|\mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2}\right), \qquad x \in \mathbf{R}.$$

Note that in this example, the location parameter μ is the mean. This is not typical.

Scale family For $\sigma > 0$, define $Y = \sigma Z$, which has PDF

$$f(y|\sigma) = \frac{1}{\sigma} f_Z\left(\frac{y}{\sigma}\right).$$

This can be thought of as a simple change of units.

Example 2.25. *Let $Z \sim \Gamma(\alpha, 1)$, where we take the shape parameter α to be fixed. Then*

$$f_Z(z) = \frac{z^{\alpha-1}}{\Gamma(\alpha)} \exp(-z), \qquad z > 0.$$

Then the random variable $Y = \sigma Z \sim \Gamma(\alpha, \sigma)$. (Be careful, though, as there are two commonly used ways of parameterizing the gamma distribution.) Its density is

$$f_Y(y|\sigma) = \frac{y^{\alpha-1}}{\sigma^\alpha \Gamma(\alpha)} \exp(-\frac{y}{\sigma}), \qquad y > 0.$$

Location-Scale family Define $W = \mu + \sigma Z$, which has PDF

$$f(w|\mu, \sigma) = \frac{1}{\sigma} f_Z\left(\frac{w-\mu}{\sigma}\right).$$

Example 2.26. *Again consider a standard normal variable $Z \sim N(0,1)$. Then the random variable $W = \mu + \sigma Z \sim N(\mu, \sigma^2)$. Its density is*

$$f_W(w|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(w-\mu)^2}{2\sigma^2}\right), \qquad w \in \mathbf{R}.$$

Again, the scale parameter σ here is the standard deviation. As can be seen from the gamma-distributed example, this is not typical.

Remark 2.27. *In statistics, we often find ourselves in the position of estimating unknown parameters from data. The parameters are often location or scale parameters of a family. Plausible experimental assumptions often determine a family of random variables that can be used to model the data, and hypotheses of scientific interest can be phrased in terms of parameters to be estimated.*

Example 2.28. *Suppose we seek to estimate the rate of decay of a radioactive substance, by measuring the times between decay events observed by a Geiger counter. From physical considerations, the inter-arrival times in seconds can be assumed to be exponentially distributed with mean β.*

$$f_T(t|\beta) = \frac{1}{\beta}\exp\left(-\frac{t}{\beta}\right), \qquad t > 0.$$

We can see that T is a scale transformation of the standard exponential random variable $Z \sim \text{Exp}(1)$, which has probability density function

$$f_Z(z) = \exp(-z), \qquad z > 0.$$

Typically, the statistical problem is to estimate β and quantify the uncertainty in the estimate that arises from using only a finite number of observations.

Proposition 2.29. Probability Integral Transform *Let $U \sim \text{Unif}[0,1]$ and let $X = F^{-1}(U)$, where F is a strictly increasing CDF. Then X is a random variable with CDF F.*

Proof. *First, note that since F is strictly increasing, F^{-1} exists. Then*

$$\begin{aligned}\Pr(X \le x) &= \Pr(F^{-1}(U) \le x) = \Pr(U \le F(x)) \\ &= F_U\left(F(x)\right).\end{aligned}$$

Now, since F is a CDF, we must have $0 \le F(x) \le 1$ for all $x \in \mathbf{R}$. Moreover, for $0 \le u \le 1$, $F_U(u) = u$. Hence

$$\Pr(X \le x) = F(x),$$

as required.

Remark 2.30. *The probability integral transform is a practically useful result. For, suppose we wish to generate a random sample $X_1, X_2, \ldots, X_n$ from some distribution whose CDF is F. So long as we can explicitly determine F^{-1}, and easily draw samples $U_1, U_2, \ldots, U_n$, our sample is just $F^{-1}(U_1), \ldots, F^{-1}(U_n)$.*

Example 2.31. *Suppose we seek a sample X from the $\text{Exp}(\beta)$ distribution from 2.28. Then*

$$F_X(x) = 1 - \exp\left(-\frac{x}{\beta}\right), \qquad x > 0.$$

It is straightforward to compute $F_X^{-1}(u) = -\beta\log(1-u)$. Using F_X^{-1} to transform U gives a sample from the exponential distribution sought.

Expectation (Review)

For a discrete random variable X, the definition of the expectation should be familiar:

$$\mathrm{E}(X) = \sum_x x \Pr(X = x),$$

where the sum is taken over the countable range of values assumed by X. A similar definition should be familiar for absolutely continuous random variables

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx.$$

As we have seen, in general random variables are neither absolutely continuous nor discrete, and it is clearly desirable to have a unified definition of expectation for all random variables. To do this properly, we would need to develop the theory of integration, but this is beyond our scope. Below is a non-examinable summary, following Grimmett and Stirzaker (5.6). Note one important point: the criterion for $\mathrm{E}(X)$ to be finite.

Abstract Expectation (non-examinable)
On a probability space $(\Omega, \mathcal{F}, \Pr)$, we first define expectation for **simple** random variables, i.e. functions $X : \Omega \to \mathbf{R}$ that take on only finitely many values. Let $\mathcal{X} = \{x_1, \ldots, x_n\}$ be the support of X, and suppose $X^{-1}(x_i) = A_i$, so that $A_1, \ldots, A_n$ forms a partition of Ω. Then X can be written in terms of indicator random variables

$$X = \sum_{i=1}^{n} x_i \mathbf{1}_{A_i},$$

and we can define expectation as

$$\mathrm{E}(X) = \sum_{i=1}^{n} x_i \Pr(A_i).$$

Any non-negative random variable $X : \Omega \to [0, \infty)$ can be written as

$$X(\omega) = \lim_{n\to\infty} X_n(\omega)$$

for an increasing sequence $(X_n)_{n\geq 1}$ of simple random variables. We then define

$$\mathrm{E}(X) = \lim_{n\to\infty} E(X_n).$$

It can be shown that this definition is independent of the increasing sequence (X_n) chosen. Note that $\mathrm{E}(X)$ may be $+\infty$.

Now, any random variable $X : \Omega \to \mathbf{R}$ can be written as $X = X^+ - X^-$, a difference of two non-negative random variables, namely

$$X^+(\omega) = \max\{0, X(\omega)\},$$

the positive part of X and

$$X^-(\omega) = -\min\{0, X(\omega)\},$$

the negative part of X.

As X^+ and X^- are non-negative random variables, $\mathrm{E}(X^+)$ and $\mathrm{E}(X^-)$ are well-defined, although possibly infinite. If at least one of these two expectations is finite, then $\mathrm{E}(X)$ is unambiguously defined by

$$\mathrm{E}(X) = \mathrm{E}(X^+) - \mathrm{E}(X^-),$$

which may be finite, $+\infty$ or $-\infty$. We see that $\mathrm{E}(X)$ is finite provided

$$\mathrm{E}|X| = \mathrm{E}(X^+) + \mathrm{E}(X^-) < \infty.$$

You may see this described in books as the condition for X to be **integrable**.

Properties of Expectation

Recal the following properties of expectation for any random variables X and Y.

1. $\mathrm{E}(aX + bY) = a\mathrm{E}(X) + b\mathrm{E}(Y)$ for all $a, b \in \mathbf{R}$.
2. If $\Pr(X \geq 0) = 1$ then $\mathrm{E}(X) \geq 0$.
3. If A is an event, then $\mathrm{E}(\mathbf{1}_A) = \Pr(A)$.

3
Multivariate Random Variables

Definition 3.1. *The joint cumulative distribution function of two random variables is given by*

$$F_{XY}(x, y) = \Pr(X \leq x, Y \leq y).$$

This definition exends in the obvious way to the joint cumulative distribution function of n random variables $X_1, \ldots, X_n$.

We will often be interested in the case where X and Y have probability densities so that

$$F_{XY}(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{XY}(s, t)\, ds\, dt.$$

The function f_{XY} is said to be a *joint probability density function* for the pair (X, Y). Such a pair of random variables is said to be **jointly absolutely continuous**.
Note that, as in the univariate case, probability density functions are not unique. We could easily change the value of f at any single point without changing the value of the integral that is the defining property of f_{XY}.

We can recover the **marginal density function** of one of the variables, say X, by integrating the joint pdf over y

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

Independence

Definition 3.2. *A* **finite collection** *of random variables $X_1, X_2, \ldots X_n$ is defined to be* **independent** *if*

$$\Pr\left(X_1 \in B_1, \ldots, X_n \in B_n\right) = \prod_{i=1}^{n} \Pr(X_i \in B_i)$$

for all Borel sets $B_1, B_2, \ldots B_n$.

An arbitrary collection $(X_i)_{i\in I}$ of random variables is independent if **every finite subcollection** *is independent.*

Proposition 3.3. *(non-examinable) Univariate functions of independent random variables are independent. Let $X_1, X_2, \ldots X_n$ be independent and $f_1, f_2, \ldots f_n$ be Borel functions, then $f_1(X_1), f_2(X_2), \ldots, f_n(X_n)$ are independent.*

Proof. *Suppose $B_1, B_2, \ldots B_n$ are arbitrary Borel sets. Then*

$$\bigcap_{i=1}^{n}\{f_i(X_i) \in B_i\} = \bigcap_{i=1}^{n}\{X_i \in f_i^{-1}(B_i)\}.$$

Since each f_i is a Borel function, each $f_i^{-1}(B_i) \in \mathcal{B}$, so that by independence,

$$\Pr\left(\{\bigcap_{i=1}^{n} f(X_i) \in B_i\}\right) = \Pr\left(\{\bigcap_{i=1}^{n} X_i \in f^{-1}(B_i)\}\right) = \prod_{i=1}^{n}\Pr\left(X_i \in f^{-1}(B_i)\right) = \prod_{i=1}^{n}\Pr\left(f(X_i) \in (B_i)\right).$$

This shows that $f_1(X_1), f_2(X_2), \ldots f_n(X_n)$ are independent whenever $X_1, X_2, \ldots X_n$ are independent.

Covariance and Correlation

Definition 3.4. *For random varibles X and Y, both with finite expectations $\mathrm{E}(X) = \mu_X$ and $\mathrm{E}(Y) = \mu_Y$, the* **covariance** *of X and Y is defined to be*

$$\mathrm{Cov}(X, Y) = \mathrm{E}\left((X - \mu_X)(Y - \mu_Y)\right).$$

The correlation between X and Y is

$$\mathrm{Cor}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

Remark 3.5. 1. *(Exercise) It is often simpler to work with the equivalent expression*

$$\mathrm{Cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y).$$

2. *Independent random variables have covariance zero, but the converse is not true in general. (See problem sheet.)*

3. *$\mathrm{Cov}(X, Y)$ has the same units as XY, so in general it does not make sense to say the covariance is large or small. The correlation however is dimensionless, indeed $-1 \leq \mathrm{Cor}(X, Y) \leq 1$.*

Proposition 3.6. *Covariance defines an* **inner product** *on an appropriately defined space.*

Proof. *To show this, there are three properties to verify. These are left as a straightforward exercise.*

1. **Bilinearity** *For random variables X,Y,Z and constants* $a, b \in \mathbf{R}$,

$$\mathrm{Cov}(aX + bY, Z) = a\mathrm{Cov}(X, Z) + b\mathrm{Cov}(Y, Z).$$

2. **Symmetry** *For any random variables* X, Y, $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$

3. **Positive semi-definiteness** *For any random variable* X,

$$\mathrm{Cov}(X, X) = \mathrm{Var}(X) \geq 0.$$

Note that the third point is not yet enough to say that Cov *is an inner product: we require positive definiteness, which does not hold in general. To surmount this problem, we define an equivalence relation* $X \sim Y$ *if and only if* $\Pr(X = Y + c) = 1$ *for some* $c \in \mathbf{R}$, *and work with the quotient space in which variables that differ by a constant with probability 1 are identified.*

Changes of variables

The data that are collected about a system do not always represent the best set of coordinates in which to analyze the system. So it will often be convenient to work with different variables from those that are first collected.

Proposition 3.7. *(Change of variables for probability densities). Suppose* $D \subseteq R^2$ *is a domain and* $T : D \to \mathbf{R}^2$ *is an invertible mapping onto* $R \subseteq \mathbf{R^2}$. *Define the* **Jacobian determinant** *of the map* T *by*

$$J(u, v) = \det \begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix},$$

and suppose that all partial derivatives exist and are continuous.
Then if $(U, V) = T(X, Y)$ *is a function of the pair of random variables* (X, Y) *with joint probability density function* f_{XY}, *a joint pdf for* (U, V) *is given by*

$$f_{UV}(u, v) = f_{XY}(x(u, v), y(u, v))|J(u, v)|.$$

Example 3.8. *The random variables* (X, Y) *have joint pdf*

$$f(x) = 18x(1 - x)y^2, \qquad 0 < x, y < 1$$

Suppose we are interested in the distribution of the product variable XY. We can determine this via the joint transformation

$$U = X, \qquad V = XY,$$

so that the mapping $T(x, y) = (x, xy)$. T is invertible with inverse $T^{-1}(u, v) = (u, v/u)$. We compute the Jacobian determinant

$$J(u, v) = \det \begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix} = \det \begin{pmatrix} 1 & 0 \\ -\frac{v}{u^2} & \frac{1}{u} \end{pmatrix} = \frac{1}{u}$$

So that the transformed density is given by

$$\begin{aligned} f_{UV}(u,v) &= f_{XY}(x(u,v),y(u,v))|J(u,v)| \\ &= 18u(1-u)\left(\frac{v}{u}\right)^2\frac{1}{u} && 0<u,\ \frac{v}{u}<1. \\ &= 18(1-u)\frac{v^2}{u^2} && 0<v<u<1 \end{aligned}$$

We can now determine the marginal density of V as

$$\begin{aligned} f_V(v) &= \int_{-\infty}^{\infty} f_{UV}(u,v)\,du = \int_v^1 18(1-u)\frac{v^2}{u^2}\,du = v^2\left[-\frac{1}{u}-\log u\right]_v^1 \\ &= v^2\left(-1+1/v+\log v\right) = v\left(1-v+v\log v\right) \qquad 0<v<1. \end{aligned}$$

Remark 3.9. *A necessary and sufficient condition for two random variables X and Y to be independent is that there exist functions $g, h : \mathbf{R} \to \mathbf{R}$ such that the joint mass or density function factorizes as*

$$f_{XY}(x,y) = g(x)h(y) \qquad \textit{for all } x,y \in \mathbf{R}.$$

Hence, we easily see that the variables X and Y of the previous example are independent. Note however, that U and V are **not** *independent, even though it appears that f_{UV} can be written as the product of terms involving u alone and v alone. This is because the support of (U,V), the range of values with non-zero probability density, is described by a non-trivial relationship between u and v. Equivalently, there are no properly defined functions g, h with domain $\mathbf{R}$ such that $f_{UV}(u,v) = g(u)h(v)$* **for all** *$u,v \in \mathbf{R}$.*

Conditioning

Recall our definition of conditional probability for events,

$$\Pr(B|A) = \frac{\Pr(B\cap A)}{\Pr(A)} \qquad \text{if } \Pr(A) > 0.$$

Suppose we take $B = \{X \le x\}$, where X is a random variable. Then we have the conditional CDF of X given A,

$$F_{X|A}(x) = \frac{\Pr(\{X\le x\}\cap A)}{\Pr(A)}.$$

As with any CDF, $F_{X|A}(x)$ completely determines the distribution of $X|A$. For a discrete random variable, we have a conditional mass function $f_{X|A}(x) = \Pr(X = x|A)$, and in the absolutely continuous case,

$$f_{X|A}(x) = \frac{d}{dx}F_{X|A}(x),$$

with

$$\Pr(X\in C|A) = \int_{x\in C} f_{X|A}(x)\,dx$$

for any $C \in \mathcal{B}$.

In statistics, we often work with the case where the joint distribution of the random variables (X, Y) is given, and we are interested in the conditional distribution of Y given that $X = x$. For cases where $\Pr(X = x) > 0$, such as when X is a discrete random variable, this is straightforward. However, in cases where $\Pr(X = x) = 0$, such as when X is a continuous random variable, we need to be more careful.

The right approach is to condition on events of positive probability of the form $\{X \in (x, x+h)\}$ for $h > 0$, and then take a limit as $h \to 0$. If f_{XY} is a joint probability density function for (X, Y), then

$$\Pr\left(Y \leq y | X \in (x, x+h)\right) = \frac{\int_x^{x+h} \int_{-\infty}^{y} f_{XY}(u, v)\, dv du}{\int_x^{x+h} f_X(u)\, du}.$$

Then we evaluate the $h \to 0$ limit by l'Hopital's rule, since both numerator and denominator tend to zero:

$$F_{Y|X}(y|x) = \lim_{h \to 0} \Pr\left(Y \leq y | X \in (x, x+h)\right) = \frac{\int_{-\infty}^{y} f_{XY}(x, v)\, dv}{f_X(x)}.$$

To spell out the differentiation of numerator and denominator, suppose G is a differentiable function such that $G'(u) = g(u)$. Then

$$\frac{d}{dh} \int_x^{x+h} g(u)\, du = \frac{d}{dh} \left(G(x+h) - G(x)\right) = g(x+h),$$

and provided g is continuous, $g(x+h) \to g(x)$ as $h \to 0$.
We then define the conditional probability density function as

$$f_{Y|X}(y|x) = \frac{d}{dy} F_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)}.$$

Remark 3.10. *Just as we think of the probability density function as satisfying*

$$f_X(x)dx = \Pr(X \in (x, x+dx)),$$

so the right interpretation of the conditional probability density is

$$f_{Y|X}(y|x)dy = \Pr(Y \in (y, y+dy) | X \in (x, x+dx)).$$

Practically, this is important since continuous measurements are always of finite precision. Mathematically, it is important so that we avoid conditioning on events of probability zero. See the problem sheet for an instance of the **Borel-Kolmogorov paradox** *, which illustrates the difficulties when attempting to condition on events of probability zero.*

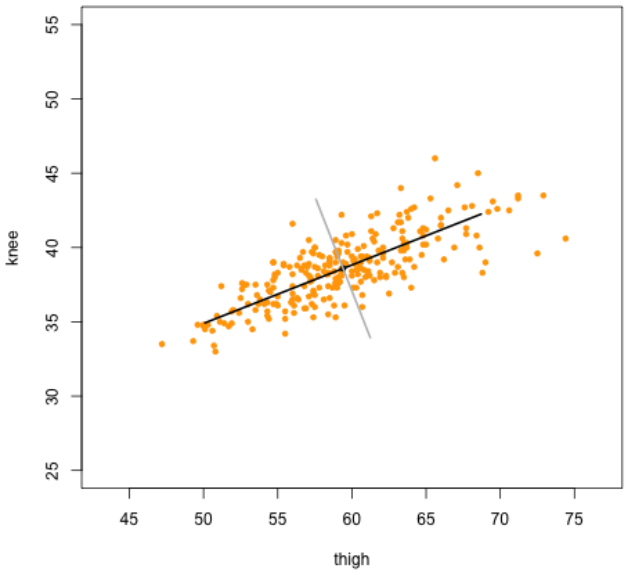


Figure 3.1: Two body measurements for a sample of 252 men. Measurements are circumference of the subjects' thigh and knee, in cm. Note the elliptical geometry of the plot, which is characteristic of the bivariate normal distribution. The principal axes are shown on the plot.

Bivariate normal distribution

For reasons that will become clear when we consider the central limit theorem, the normal distribution is commonly encountered in statistics. Data will often take the form of d different measurements on n different experimental sujects. Across subjects, measurements will often be correlated. To allow arbitrary correlations between such measurements, we introduce the **multivariate normal** distribution, begnning with the simpler case $d = 2$.

The standard bivariate normal distribution . We will begin by considering the standardized bivariate normal distribution, whose probability density function is

$$f(x, y|\rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right) \qquad (x, y) \in \mathbf{R}^2,$$

for $-1 < \rho < 1$. Here, standardization means that we choose the origin and units of measurement so that, marginally, we have $\mathrm{E}(X) = \mathrm{E}(Y) = 0$ and $\mathrm{Var}(X) = \mathrm{Var}(Y) = 1$. Since normal variables form a location-scale family, there is no loss of generality here. Note though that when working with data, the mean and variance are **parameters** that we may need to estimate.

It is often useful in working with the bivariate normal to complete

the square in the quadratic term inside the exponent.

$$x^2 - 2\rho xy + y^2 = (x - \rho y)^2 + (1 - \rho^2)y^2.$$

This makes it straightforward to evaluate integrals involving the bivariate joint density. E.g. suppose we wish to evaluate the marginal density of Y. This is

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f_{XY}(x,y)\,dx = \int_{-\infty}^{\infty} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}(x-\rho y)^2 - \frac{1}{2}y^2\right) dx \\ &= \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}y^2) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}(x-\rho y)^2\right) dx \\ &= \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}y^2), \end{aligned}$$

where the final equality follows because the integrand is the probability density function of a $N(\rho y, 1-\rho^2)$ variable. We see that Y, and therefore also, by symmetry, X, have standard normal marginal distributions, $X, Y \sim N(0,1)$.

The rewriting above corresponds to writing the joint density as

$$f_{XY}(x,y) = f_{X|Y}(x|y) f_Y(y).$$

We can read off the conditional density $f_{X|Y}$ from the joint density, just by noting the functional dependence of the joint density on x. We see that $f_{X|Y}$ and f_{XY} have the same dependence on x, and differ only by terms that depend on y. In the conditional density for x, we regard y as fixed, and so the terms depending on y are simply part of the normalizing constant. Since we know that this conditional density integrates to 1, the constant of proportionality can be determined.

Hence we see that

$$f_{X|Y}(x|y) \propto \exp\left(-\frac{1}{2(1-\rho^2)}(x-\rho y)^2 - \frac{1}{2}y^2\right).$$

This density has the same functional dependence on x as has a $N(\rho y, 1-\rho^2)$ variable. Hence $X|Y = y \sim N(\rho y, 1-\rho^2)$.

Note therefore that X and Y are not independent in general. We now calculate their covariance. Firstly,

$$\begin{aligned}
\mathrm{E}(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_{X|Y}(x|y) f_Y(y)\,dx\,dy \\
&= \int_{-\infty}^{\infty} y f_Y(y) \int_{-\infty}^{\infty} x f_{X|Y}(x|y)\,dx\,dy \\
&= \int_{-\infty}^{\infty} y f_Y(y) \mathrm{E}(X|Y)\,dy \\
&= \int_{-\infty}^{\infty} y f_Y(y)\,\rho y\,dy = \rho \mathrm{E}(Y^2) = \rho.
\end{aligned}$$

Remark 3.11. *Really, this is just the law of iterated expectation:*

$$\mathrm{E}(XY) = \mathrm{E}_Y\left(\mathrm{E}(XY|Y=y)\right) = \mathrm{E}_Y\left(\rho Y\, Y\right) = \rho \mathrm{E}_Y\left(Y^2\right) = \rho.$$

Since now $\mathrm{E}(X) = \mathrm{E}(Y) = 0$, this gives

$$\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) = \rho.$$

Remark 3.12. *Note that the conditional distribution of X given $Y = y$ is Normal with a mean ρy, which is a function of y, but a constant variance $1-\rho^2$.*

Remark 3.13. *The bivariate normal distribution defined represents variables that have been standardized - measured such that they have mean 0 and variance 1. The most general form of the bivariate normal density allows the X and Y variables to have arbitrary mean and arbitrary (positive) variance:*

$$f_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\mu},\Sigma) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right),$$

where $\boldsymbol{\mu} = (\mu_X, \mu_Y)$ is the mean vector and

$$\Sigma = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$$

is the variance-covariance matrix of the random vector $\boldsymbol{x} = (x, y)$.

Indeed, it is more natural to write the density in terms of the vector $\boldsymbol{x} = (x, y)$ as

$$f_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\mu},\Sigma) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^t\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right).$$

From this representation, we can see quite clearly that the probability density depends on $\boldsymbol{x}$ through a positive definite quadratic form in $\boldsymbol{x}$. This means that contours of equal probability density take the form of ellipses, as can be seen in Figure 3.2. So then the bivariate normal density is of the form

$\exp(-d(\boldsymbol{x}, \boldsymbol{\mu})^2)$ *for a generalized distance function* d*. In statistical work, this distance function is called* **Mahalanobis distance**.

When working with data, the properties of this distance function have an intuitive explanation in terms of the relationship between the variables. Suppose we consider an indidivual with a thigh circumference of $\sim$ 70cm*. How unusual is such an individual? This is quantified by the individual's distance from the mean point* $(59, 39)$*. But the right notion of distance is clearly not just given by Euclidean distance in Figure 3.1: the given thigh measurement is much more likely for individuals with above-average knee measurements, because the two variables are positively correlated. Instead, the right notion of distance takes into account the fact that the measured variables are correlated. Mahalanobis distance is equivalent to using Euclidean distance on the transformed variables shown as principal axes in 3.1, because these variables are uncorrelated. Figure 3.2 shows contour lines connecting points of equal probability density for the data shown in 3.1.*

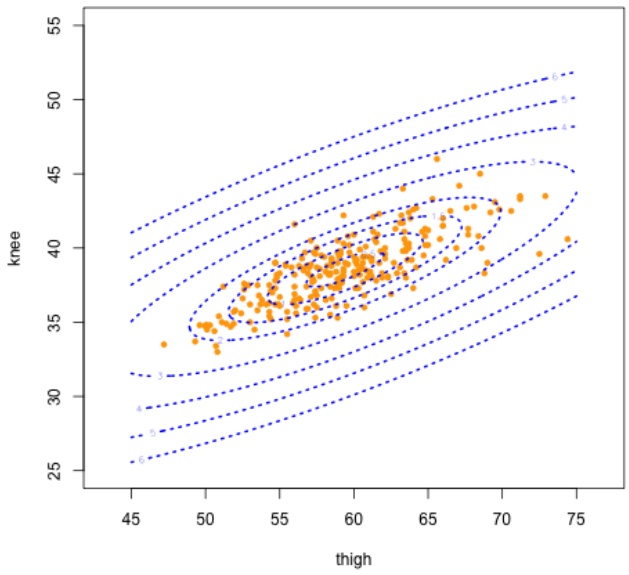


Figure 3.2: Contours of bivariate normal probability density for the data in Figure 3.1.

Multivariate normal distribution

The idea of the bivariate normal distribution readily extends to a d-dimensional vector. The general multivariate normal density is specified in terms of its mean vector $\boldsymbol{\mu}$ and positive definite variance-

covariance matrix Σ as

$$f_X(\boldsymbol{x}|\boldsymbol{\mu},\Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}}\,(\det\Sigma)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^t\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \qquad \boldsymbol{x}\in\mathbf{R}^d.$$

Definition 3.14. *For a random vector $\boldsymbol{X} = (X_1,\ldots,X_d)$, the* **variance-covariance** *matrix is the $d\times d$ matrix with (i,j)-th entry $\mathrm{Cov}(X_i,X_j)$.*

Remark 3.15.

- *By construction, a variance-covariance matrix is always symmetric, because $\mathrm{Cov}(X_i,X_j) = \mathrm{Cov}(X_j,X_i)$.*

- *The ith diagonal entry is $\mathrm{Cov}(X_i,X_i) = \mathrm{Var}(X_i)$.*

- *For any constant vector $\boldsymbol{a}\in\mathbf{R}^d$,*

  $$\mathrm{Var}(\boldsymbol{a}^t\boldsymbol{X}) = \mathrm{Var}(\sum_{i=1}^d a_iX_i) = \sum_{i=1}^d a_i^2\mathrm{Var}(X_i) + 2\sum_{i<j} a_ia_j\mathrm{Cov}(X_i,X_j) = \boldsymbol{a}^t\Sigma\boldsymbol{a}.$$

  Since variances are always non-negative, it follows that $\boldsymbol{a}^t\Sigma\boldsymbol{a}\geq 0$ for all vectors $\boldsymbol{a}\in\mathbf{R}^d$. This says that Σ is positive semi-definite.

  If, in fact, $\mathrm{Var}(\boldsymbol{a}^t\boldsymbol{X}) = 0$ for some $\boldsymbol{a}\neq 0$, then there is a linear combination of the entries of $\boldsymbol{X}$ that is constant with probability 1. This is a somewhat degenerate case, corresponding to a singular distribution: there is no d-dimensional probability density. Equivalently, it says that the variance-covariance matrix Σ is not invertible. In the bivariate case, it corresponds to a correlation $\rho\in\{\pm 1\}$, which we specifically excluded. We will consider only the case where Σ is positive definite.

Proposition 3.16. *If $\boldsymbol{X}\sim MVN_d(\boldsymbol{\mu},\Sigma)$ and A is an invertible $d\times d$ matrix, then the random variable $Y = AX\sim MVN_d(A\mu, A\Sigma A^t)$.*

Proof. *By the theorem on invertible multivariate transformations in the general $d\times d$ case, if $\boldsymbol{Y} = A\boldsymbol{X}$ then $X = A^{-1}Y$ so that*

$$\frac{\partial x_i}{\partial y_j} = \left(A^{-1}\right)_{ij},$$

and so the Jacobian determinant is just $\det A^{-1}$.

$$
\begin{aligned}
f_Y(\boldsymbol{y}) &= f_X(A^{-1}\boldsymbol{y})|\det A^{-1}| \\
&= \frac{1}{(2\pi)^{\frac{d}{2}}(\det\Sigma)^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(A^{-1}\boldsymbol{y}-\boldsymbol{\mu})^t\Sigma^{-1}(A^{-1}\boldsymbol{y}-\boldsymbol{\mu})\right)|\det A^{-1}| \\
&= \frac{1}{(2\pi)^{\frac{d}{2}}(\det\Sigma)^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\boldsymbol{y}-A\boldsymbol{\mu})^t\left(A^{-1}\right)^t\Sigma^{-1}A^{-1}(\boldsymbol{y}-A\boldsymbol{\mu})\right)|\det A^{-1}| \\
&= \frac{1}{(2\pi)^{\frac{d}{2}}(\det\Sigma)^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\boldsymbol{y}-A\boldsymbol{\mu})^t\left(A\Sigma A^t\right)^{-1}(\boldsymbol{y}-A\boldsymbol{\mu})\right)|\det A^{-1}|,
\end{aligned}
$$

using the standard results that for matrices P and Q, $(PQ)^{-1} = Q^{-1}P^{-1}$ and $\left(P^{-1}\right)^t = \left(P^t\right)^{-1}$.

Noting now the standard properties of the determinant:

$$\det PQ = \det P \det Q, \qquad \det P = \det P^t,$$

we see that $\det A^{-1} = 1/\det A$ *and* $|\det A| = \left(\det AA^t\right)^{\frac{1}{2}}$*, so that*

$$
\begin{aligned}
f_Y(\boldsymbol{y}) &= f_X(A^{-1}\boldsymbol{y})|\det A^{-1}| \\
&= \frac{1}{(2\pi)^{\frac{d}{2}}(\det AA^t)^{\frac{1}{2}}(\det\Sigma)^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\boldsymbol{y}-A\boldsymbol{\mu})^t\left(A\Sigma A^t\right)^{-1}(\boldsymbol{y}-A\boldsymbol{\mu})\right) \\
&= \frac{1}{(2\pi)^{\frac{d}{2}}(\det A\Sigma A^t)^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\boldsymbol{y}-A\boldsymbol{\mu})^t\left(A\Sigma A^t\right)^{-1}(\boldsymbol{y}-A\boldsymbol{\mu})\right),
\end{aligned}
$$

which shows that $Y \sim MVN_d(A\boldsymbol{\mu}, A\Sigma A^t)$*.*

Proposition 3.17. *We can always find a linear transformation Q of the multivariate normal vector $\boldsymbol{X} = (X_1, \ldots, X_n)^t$ such that the entries of $\boldsymbol{Z} = Q\boldsymbol{X}$ are uncorrelated, and indeed independent, random variables.*

Proof *As Σ is a symmetric, positive definite matrix, there exists an orthogonal matrix Q, i.e. with $QQ^t = Q^tQ = I_d$, such that*

$$Q\Sigma Q^t = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{pmatrix},$$

where $\lambda_1, \ldots \lambda_d$ are the eigenvalues of A, which are real and positive. Then define $\boldsymbol{z} = Q^t(\boldsymbol{x}-\boldsymbol{\mu})$, so that $Q\boldsymbol{z} = \boldsymbol{x}-\boldsymbol{\mu}$ and note that

$$(Q^t\Sigma Q)^{-1} = Q^{-1}\Sigma^{-1}\left(Q^t\right)^{-1} = Q^t\Sigma^{-1}Q = \begin{pmatrix} \frac{1}{\lambda_1} & & \\ & \ddots & \\ & & \frac{1}{\lambda_d} \end{pmatrix}.$$

Then we have that

$$(\boldsymbol{x}-\boldsymbol{\mu})^t\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = \boldsymbol{z}^t Q^t \Sigma^{-1} Q\boldsymbol{z} = \sum_{i=1}^{d} \frac{z_i^2}{\lambda_i},$$

so, noting that the absolute value of the Jacobian determinant of the transformation is 1*, since* Q *is orthogonal,*

$$\begin{aligned} f_{\boldsymbol{Z}}(\boldsymbol{z}) &= f_X(\boldsymbol{\mu} + Q\boldsymbol{z}) \\ &= \frac{1}{(2\pi)^{\frac{d}{2}} (\det\Sigma)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\sum_{i=1}^{d}\frac{z_i^2}{\lambda_i}\right) \\ &= \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi\lambda_i}} \exp\left(-\frac{z_i^2}{2\lambda_i}\right), \end{aligned}$$

where in the last equality we have used the fact that $\det\Sigma = \prod_{i=1^d} \lambda_i$.

We seen then that $Z_1, \ldots Z_d$ *are independent normal random variables, because the joint density factorizes over all of* $\mathbf{R}^d$*. The variables are shown in Figure 3.1 as the orthogonal principal directions in the ellipse. For the three variables in 3.3, there are are three orthogonal principal directions.*

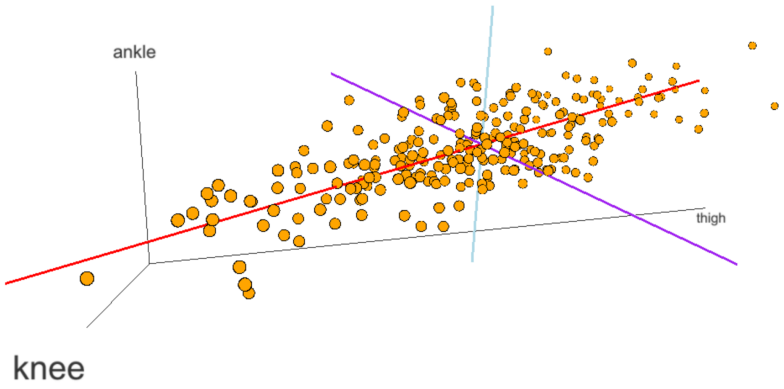


Figure 3.3: Three body measurements for 252 men. Measurements are circumference of the subjects' thigh, ankle and knee. The red, purple and blue lines show the three orthogonal principal axes of the sample variance-covariance matrix.

Order statistics

Data can often be assumed to be independent, identical draws from some continuous distribution. Suppose we have such a random sample $X_1, \ldots X_n$, drawn from the distribution with cumulative distribution function F_X and probability density function f_X. It is common to report summaries of the data that relate to the ordering

of the sample, such as $\min X_i$, $\max X_i$ and the sample median, which is either the middle ordered value (for n odd) or the average of the two middle values (for n even).

This suggests that we should investigate the joint distribution of *ordered* samples from a distribution. We let Y_1 be the smallest amongst sampled variables, Y_2 the next smallest etc. Since we assume the random variable has a continuous distribution, we neglect the possibility of ties. The notation $X_{(k)}$ for Y_k is common in statistics. $(Y_1, \ldots, Y_n)$ is the vector of **order statistics** of the random vector $(X_1, X_2, \ldots X_n)$.

By symmetry, and independence, the joint density of the order statistics is

$$f(y_1, \ldots, y_n) = \begin{cases} n! \prod_{i=1}^n f_X(y_i), & y_1 < y_2 < \ldots < y_n \\ 0 & \text{otherwise.} \end{cases}$$

The marginal density of Y_k, for $k \in \{1, \ldots n\}$ is given by

$$f_k(y) = k \binom{n}{k} f_X(y) F_X(y)^{k-1} \left(1 - F_X(y)\right)^{n-k}.$$

To derive the marginal density, note that the event $\{Y_k \leq y\}$ occurs if and only if the event $\{N_y \geq k\}$ occurs, where N_y counts the number of the X_i that are at most y.
Since $N_y \sim \text{Bin}(n, F_X(y))$, we can determine the CDF of Y_k as

$$F_k(y) = \Pr(N_y \geq k) = \sum_{j=k}^{n} \binom{n}{j} F_X(y)^j \left(1 - F_X(y)\right)^{n-j}.$$

The density now follows on differentiating this expression. (Exercise).

Example 3.18. *Maximum and minimum*
These two particular cases of the general order statistics result are often useful, and easy to interpret. The minimum of n values is greater than y if

and only if they are all greater than y, so that we have the equality of events

$$\{Y_1 > y\} = \bigcap_{i=1}^{n} \{X_i > y\},$$

so that by independence of the X_i,

$$\Pr(Y_1 \leq y) = 1 - \Pr(X > y)^n = 1 - (1 - F_X(y))^n.$$

The maximum of n values is smaller than y if and only if they are all smaller

than y, so we have a similar equality of events

$$\{Y_n \leq y\} = \bigcap_{i=1}^{n} \{X_i \leq y\},$$

so that again by independence,

$$\Pr(Y_n \leq y) = \Pr(X \leq y)^n = (F_X(y))^n.$$

Example 3.19. *Suppose $X_1, X_2, \ldots, X_n \sim \text{Exp}(\lambda)$ are a random sample, and λ is the rate parameter. Then for each i,*

$$F_{X_i}(x) = 1 - \exp(-\lambda x), \qquad x > 0,$$

so $Y_1 = \min X_i$ has distribution

$$\Pr(Y_1 \leq y) = 1 - \Pr(Y_1 > y) = 1 - \Pr(X_i > y)^n = 1 - \exp(-n\lambda y), \qquad y > 0.$$

Hence $Y_1 \sim \text{Exp}(n\lambda)$. This is often a useful result.

4
Convergence of Random Variables

Motivation

In statistics, we are often interested in evaluating the uncertainty associated with an estimate. For example, we might be interested in evaluating the prevalence p of a disease in a population. To do this, we might take a random sample of n individuals and estimate p by the proportion of individuals in the sample who have the disease. More formally, if $X_1, \dots X_n$ are independent Bernoulli random variables with success probability p, then the maximum likelihood estimator of p is

$$\widehat{p} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then by linearity

$$\mathrm{E}(\widehat{p}) = \frac{1}{n}\mathrm{E}\left((\sum_{i=1}^{n} X_i\right) = p$$

and since the observations in a random sample are independent,

$$\mathrm{Var}\,(\widehat{p}) = \frac{1}{n^2}\mathrm{Var}(\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{p(1-p)}{n}.$$

Note then that our estimator has desirable properties. As n becomes large, the distribution of the estimator becomes concentrated around its expectation, which is the true value, with ever smaller variance. We can imagine taking a sequence $(\widehat{p}_n)_{n\geq 1}$ with the nth element of the sequence based on a sample of size n. Note that each element of the sequence is a random variable.

The aim of this chapter is to build up a framework for talking about the convergence properties of such sequences. As random variables are functions, there are several different senses in which they can be said to converge. As well as the example discussed above, where n

represented a notional sample size, questions of convergence also frequently arise for stochastic processes, the subject of the final chapter. In this setting, n is most usually interpreted as a discrete time variable.

Convergence in probability

Definition 4.1. *A sequence $X_1, X_2, \ldots$ of random variables is said to* **converge in probability** *to a random variable X, written $X_n \xrightarrow{P} X$ if for all $\epsilon > 0$,*

$$\lim_{n\to\infty} \Pr\left(|X_n - X| \geq \epsilon\right) = 0.$$

Example 4.2. *The random variables in the sequence $(X_n)_{n\geq 1}$ have probability distribution*

$$\Pr(X_n = k) = \begin{cases} 1 - \frac{1}{n} & k = 0 \\ \frac{1}{n} & k = n. \end{cases}$$

Then $X_n \xrightarrow{P} 0$, since for any $\epsilon > 0$

$$\Pr(|X_n| \geq \epsilon) = \Pr(X_n = n) \to 0.$$

Note here that X_n converges in probability to a degenerate random variable, i.e. the constant $X = 0$.

Example 4.3. *Let $U_1, U_2, \ldots$ be a sequence of independent* $\text{Unif}(0, 1)$ *random variables. For each $n \geq 1$, define*

$$M_n = \max_{1\leq i\leq n} U_i.$$

Then $M_n \xrightarrow{P} 1$.

Proof *Note that $M_n \leq 1$ so that $|M_n - 1| = 1 - M_n$. Let $\epsilon > 0$*

$$\Pr\left(1 - M_n \geq \epsilon\right) = \Pr(M_n \leq 1 - \epsilon).$$

Recalling the previous section on order statistics, we see that for $0 < \epsilon < 1$,

$$\Pr(M_n \leq 1 - \epsilon) = \prod_{i=1}^{n} \Pr(U_i \leq 1 - \epsilon) = (1 - \epsilon)^n.$$

Then $\Pr(M_n \leq 1 - \epsilon) \to 0$ as $n \to \infty$ and so $M_n \xrightarrow{P} 1$.

We will return to this example to study the fluctuations around the limit once we have developed the concept of convergence in distribution.

Proposition 4.4. *Markov's inequality. Let X be a random variable taking only non-negative values and let $a > 0$ be a constant. Then*

$$\Pr(X \geq a) \leq \frac{\mathrm{E}(X)}{a}.$$

Proof. *Let $A = [a, \infty)$ and define the indicator*

$$I_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$$

Certainly $X \geq aI_A(X)$, so that

$$\mathrm{E}(X) \geq a\mathrm{E}(I_A(X)) = a\Pr(X \geq a).$$

The result follows.

We apply Markov's inequality to deduce Chebychev's inequality.

Proposition 4.5. *Let X be a random variable with finite mean $\mathrm{E}(X) = \mu$ and finite variance $\mathrm{Var}(X) = \sigma^2$. Then for any $\epsilon > 0$,*

$$\Pr(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}.$$

Proof. *Define the non-negative random variable $Y = (X - \mu)^2$.*

$$\Pr(|X - \mu| \geq \epsilon) = \Pr((X - \mu)^2 \geq \epsilon^2) = \Pr(Y \geq \epsilon^2).$$

Applying Markov's inequality to Y with $a = \epsilon^2$ gives

$$\Pr(Y \geq \epsilon^2) \leq \frac{\mathrm{E}(X - \mu)^2}{\epsilon^2} = \frac{\sigma^2}{\epsilon^2}.$$

Definition 4.6. *For a sequence of random variables $X_1, X_2, \ldots$, we define*

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_n,$$

the **sample mean** *of the first n variables.*

Proposition 4.7. Weak Law of Large Numbers. *Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables with finite mean μ and finite variance σ^2. Then $\bar{X}_n \xrightarrow{P} \mu$.*

Proof. *First note that, $\mathrm{E}(\bar{X}_n) = \mu$, by linearity of expectation. Then using the independence of the X_i,*

$$\mathrm{Var}(\bar{X}_n) = \mathrm{Var}\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{\sigma^2}{n}.$$

We now apply Chebychev's inequality to see that for any $\epsilon > 0$,

$$\Pr\left(|\bar{X}_n - \mu| \geq \epsilon\right) \leq \frac{\mathrm{Var}(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}.$$

The right hand side clearly tends to zero as $n \to \infty$, and so we have established $\bar{X}_n \xrightarrow{P} \mu$.

Remark 4.8. *Note that $\mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$ holds provided the variables X_i are uncorrelated, a strictly weaker condition than the independence we have assumed.*

Convergence in distribution

Definition 4.9. *A sequence of random variables $X_1, X_2, \ldots$ with CDFs $F_1, F_2, \ldots$ is said to* **converge in distribution** *to a random variable X with CDF F_X, written $X_n \xrightarrow{D} X$ if*

$$\lim_{n\to\infty} F_n(x) = F_X(x)$$

at all points $x \in \mathbf{R}$ for which F_X is continuous.

Example 4.10. *We saw that the maximum of n independent uniform variables, $M_n \xrightarrow{P} 1$. We will now examine the random variable describing the fluctuations around this limit.*

Let $Y_n = n(1 - M_n)$ be the scaled difference between M_n and its limit. We scale up by n to 'zoom in', since we know the difference between M_n and 1 diminishes as $n \to \infty$. Then for $y \in (0, n)$,

$$\Pr(Y_n \leq y) = \Pr\left(1 - M_n \leq \frac{y}{n}\right) = \Pr\left(M_n \geq 1 - \frac{y}{n}\right) = 1 - \left(1 - \frac{y}{n}\right)^n.$$

Hence we see that for any $y > 0$,

$$\lim_{n\to\infty} \Pr(Y_n \leq y) = 1 - \lim_{n\to\infty} \left(1 - \frac{y}{n}\right)^n = 1 - \exp(-y).$$

It follows that $Y_n \xrightarrow{D} Y$, where $Y \sim \text{Exp}(1)$

Example 4.11. *Let $X_n \sim \text{Unif}(-\frac{1}{n}, \frac{1}{n})$, with CDF*

$$F_n(x) = \begin{cases} 0 & x < -\frac{1}{n} \\ \frac{nx+1}{2} & -\frac{1}{n} \leq x < \frac{1}{n} \\ 1 & x \geq \frac{1}{n}. \end{cases}$$

For each $x < 0$, clearly $F_n(x) \to 0$ and for each $x > 0$, $F_n(x) \to 1$. Hence we see that $X_n \xrightarrow{D} X$, the constant random variable taking the value 0 with probability 1.

Note however that $F_n(0) = \frac{1}{2}$ for all $n \geq 1$, although $F_X(0) = 1$. This is consistent with the definition of convergence in distribution, because F_X is not (left) continuous at 0: $F_X(h) = 0$ for any $h < 0$.

Proposition 4.12. *Convergence in probability implies convergence in distribution.*

Proof. *Let $X_1, X_2, \ldots$ be a sequence of random variables with CDFs $F_1, F_2, \ldots$. Suppose that $X_n \xrightarrow{P} X$, a random variable with CDF F. We will show that $F_n(x) \to F(x)$ at all continuity points of F.*

Let x be any continuity point of F, and let $\epsilon > 0$. We begin by observing that if $X_n \leq x$, then either $X \leq x + \epsilon$, or X_n and X are separated by more than ϵ. Hence

$$\{X_n \leq x\} \subseteq \{X \leq x + \epsilon\} \cup \{|X_n - X| > \epsilon\}.$$

To see this formally, note that if $X_n \leq x$ but $|X_n - X| \leq \epsilon$ then

$$-\epsilon \leq X_n - X \leq \epsilon.$$

so that by considering the left-hand inequality,

$$X \leq X_n + \epsilon \leq x + \epsilon.$$

Then by a union bound, we must have

$$F_n(x) = \Pr(X_n \leq x) \leq \Pr(X \leq x + \epsilon) + \Pr(|X_n - X| > \epsilon).$$

A similar argument yields a lower bound for $F_n(x)$: if $X \leq x - \epsilon$, then either $X_n \leq x$ or X_n and X are separated by more than ϵ, so

$$\Pr(X \leq x - \epsilon) \leq \Pr(X_n \leq x) + \Pr(|X_n - X| > \epsilon).$$

Combining these two inequalities gives a two-sided bound for $F_n(x)$:

$$\Pr(X \leq x - \epsilon) - \Pr(|X_n - X| > \epsilon) \leq \Pr(X_n \leq x) \leq \Pr(X \leq x + \epsilon) + \Pr(|X_n - X| > \epsilon).$$

Passing to the limit $n \to \infty$, the term $\Pr(|X_n - X| > \epsilon) \to 0$, since $X_n \xrightarrow{P} X$, so that

$$F_X(x - \epsilon) \leq \lim_{n \to \infty} \Pr(X_n \leq x) \leq F_X(x + \epsilon).$$

Noting now that F_X is continuous at x, and that $\epsilon > 0$ is arbitrary, we see that $F_n(x) \to F_X(x)$ as $n \to \infty$, and so $X_n \xrightarrow{D} X$.

Convergence in distribution is strictly weaker than convergence in probability, as the following example shows.

Example 4.13. *Let $U \sim \text{Unif}(-1, 1)$ and for $n \geq 1$, define $U_n = -U$. Then U and U_n are identically distributed, so trivially $U_n \xrightarrow{D} U$. However, taking $\epsilon = \frac{1}{2}$,*

$$\Pr\left(|U_n - U| \geq \frac{1}{2}\right) = \Pr\left(|2U| \geq \frac{1}{2}\right) = \frac{3}{4}.$$

While convergence in distribution is weaker in general than convergence in probability, they are equivalent in an important special case. We are often interested in cases where $X_n \xrightarrow{D} c$, where $c \in \mathbf{R}$ is a constant. In this case, we do indeed have that $X_n \xrightarrow{P} c$.

Proposition 4.14. *Suppose $(X_n)_{n\geq 1}$ is a sequence of random variables such that $X_n \xrightarrow{D} c$ for $c \in \mathbf{R}$. Then $X_n \xrightarrow{P} c$.*

Proof. The constant random variable $X = c$ has as its cdf the point mass at c:

$$F(x) = \begin{cases} 0 & x < c \\ 1 & x \geq c. \end{cases}$$

Let F_n be the cdf of X_n. Since F is continuous at all points except c, we see that for any $\epsilon > 0$,

$$\lim_{n\to\infty} F_n(c-\epsilon) = 0, \qquad \lim_{n\to\infty} F_n(c+\epsilon) = 1.$$

To demonstrate convergence in probability, note that we have the bound

$$\begin{aligned} \Pr(|X_n - c| \geq \epsilon) &= \Pr(X_n \leq c-\epsilon) + \Pr(X_n \geq c+\epsilon) \\ &\leq \Pr(X_n \leq c-\epsilon) + \Pr\left(X_n > c + \frac{\epsilon}{2}\right) \\ &= F_n(c-\epsilon) + 1 - F_n\left(c+\frac{\epsilon}{2}\right). \end{aligned}$$

So that when $n \to \infty$,

$$0 \leq \lim_{n\to\infty} \Pr(|X_n - c| \geq \epsilon) \leq \lim_{n\to\infty} F_n(c-\epsilon) + 1 - F_n\left(c+\frac{\epsilon}{2}\right) = 0,$$

which establishes the result.

Limit events

If $(\Omega, \mathcal{F}, \Pr)$ is a probability space, and $A_1, A_2, \ldots$ is a sequence of events, then we have seen that $\mathcal{F}$ contains events such as $\bigcup_{n=1}^{\infty} A_n$, the event that at least one of the A_n occur, and $\bigcap_{n=1}^{\infty} A_n$, the event that all of the A_n occur.

We are often interested in understanding whether or not infinitely many of the A_n occur, often written $\{A_n \text{ i.o.}\}$, for A_n *infinitely often*. Closely related is the event that all but finitely many of the A_n occur, written $\{A_n \text{ a.a.}\}$, for A_n *almost always*. Clearly $\{A_n \text{ a.a.}\} \subseteq \{A_n \text{ i.o.}\}$. To study these events, we need to introduce a new concept. As a motivation, we recall the lim sup and lim inf of a real sequence (a_n). For each $n \geq 1$, define the sequences

$$b_n = \inf_{m\geq n} a_m, \qquad c_n = \sup_{m\geq n} a_m.$$

Note that (b_n) is an increasing sequence and (c_n) is a decreasing sequence, as depicted in Figure 4.1. As these sequences are monotonic,

they either tend to a limit or to $\pm\infty$. Hence we can unambiguously define

$$\liminf_{n\to\infty} a_n = \lim_{n\to\infty} b_n, \qquad \limsup_{n\to\infty} a_n = \lim_{n\to\infty} c_n.$$

Moreover, because b_n and c_n are monotonic, we see that these limits are in fact given by

$$\liminf_{n\to\infty} a_n = \lim_{n\to\infty} b_n = \sup_{n\geq 1} b_n = \sup_{n\geq 1} \inf_{m\geq n} a_m$$

and

$$\limsup_{n\to\infty} a_n = \lim_{n\to\infty} c_n = \inf_{n\geq 1} c_n = \inf_{n\geq 1} \sup_{m\geq n} a_m.$$

You may recall from analysis that (a_n) converges if and only if $\liminf_{n\to\infty} a_n = \limsup_{n\to\infty} a_n$.

We now consider an analogous construction for sets, in which the $\leq$ relation for real numbers corresponds to the inclusion relation, $\subseteq$. Suppose we have a sequence of sets (A_n). We can manufacture an increasing and a decreasing sequence of sets from A_n by

$$B_n = \bigcap_{m=n}^{\infty} A_m, \qquad C_n = \bigcup_{m=n}^{\infty} A_m.$$

Note how $\cap$ and $\cup$ act as lower and upper bounds, respectively. We then define

$$\liminf_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} \bigcap_{m\geq n} A_m, \qquad \limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{m\geq n} A_m.$$

Our next result gives the probabilistic interpretation of these limit events.

Proposition 4.15.

$$\{A_n \text{ i.o.}\} = \limsup_{N\to\infty} A_N = \bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} A_n \in \mathcal{F},$$
$$\{A_n \text{ a.a.}\} = \liminf_{N\to\infty} A_N = \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} A_n \in \mathcal{F}.$$

Proof. *$\omega \in \Omega$ lies in infinitely many of the events A_n if and only if, for any $N \in \mathbf{N}$, $\omega \in A_n$ for some $n \geq N$. So for each $N \in \mathbf{N}$, $\omega \in \bigcup_{n=N}^{\infty} A_n$. This then says that*

$$\{A_n \text{ i.o.}\} = \bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} A_n.$$

Since we have written $\{A_n \text{ i.o.}\}$ as a countable intersection of countable unions of events, $\{A_n \text{ i.o.}\} \in \mathcal{F}$.

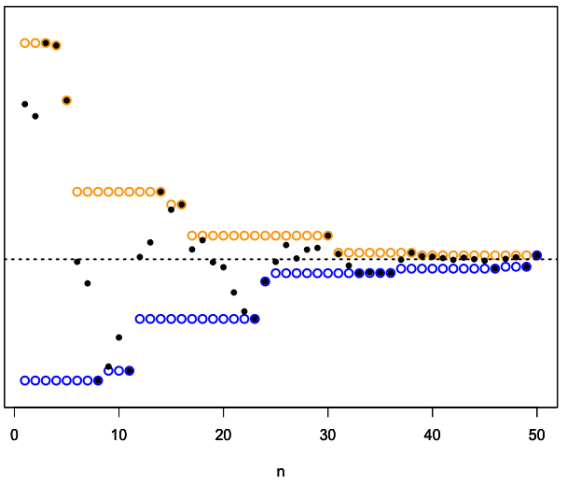


Figure 4.1: For the sequence (a_n) in black, the sequence $b_n = \inf_{m \geq n} a_m$ is shown in blue and the sequence $c_n = \sup_{m \geq n} a_m$ is shown in orange. Note that both sequences are monotonic. The dotted line is the limit of the sequence. The original sequence a_n has a limit if and only if b_n and c_n approach the same limit.

Similarly, $\omega \in \Omega$ lies in all but finitely many of the events A_n if and only if there exists $N \in \mathbf{N}$ such that for all $n \geq N$, $\omega \in A_n$. For such an N, we then have $\omega \in \bigcap_{n=N}^{\infty} A_n$. This then says that

$$\{A_n \text{ a.a.}\} = \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} A_n.$$

We have expressed $\{A_n \text{ a.a.}\}$ as a countable union of countable intersections of events, so $\{A_n \text{ a.a.}\} \in \mathcal{F}$

Remark 4.16. *The complement of $\{A_n \text{ i.o.}\}$ is the event that only finitely many of the A_n occur. By de Morgan's law, this is*

$$\{A_n \text{ i.o.}\}^c = \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} A_n^c.$$

This says that $\omega \in \{A_n \text{ i.o.}\}^c$ if and only if there exists $N \in \mathbf{N}$ such that, for all $n \geq N$, $\omega \in A_n^c$.
Hence

$$\{A_n \text{ i.o.}\}^c = \{A_n^c \text{ a.a.}\}.$$

Borel-Cantelli lemmas

Proposition 4.17. *Let $(\Omega, \mathcal{F}, \Pr)$ be a probability space and let $A_1, A_2, \ldots$ be a sequence of events. Then*

1. *if $\sum_{n=1}^{\infty} \Pr(A_n) < \infty$ then $\Pr(\{A_n \text{ i.o.}\}) = 0$.*
2. *if $\sum_{n=1}^{\infty} \Pr(A_n) = \infty$ and $A_1, A_2, \ldots$ are independent events, then $\Pr(\{A_n \text{ i.o.}\}) = 1$.*

Proof.

1. Define $B_N = \bigcup_{n=N}^{\infty} A_n$. Then $B_1 \supseteq B_2 \supseteq \ldots$ is a decreasing sequence of sets. We see that by continuity applied to the decreasing sequence (B_N),

$$\Pr(\{A_n \text{ i.o.}\}) = \Pr\left(\bigcap_{N=1}^{\infty} B_N\right) = \lim_{N\to\infty} \Pr(B_N).$$

   But

$$0 \leq \Pr(B_N) \leq \sum_{n=N}^{\infty} \Pr(A_n) \to 0,$$

   since $\sum_{n=1}^{\infty} \Pr(A_n) < \infty$.

2.

$$\Pr(\{A_n \text{ i.o.}\}^c) = \Pr\left(\{A_n^c \text{ a.a.}\}\right) = \Pr\left(\bigcup_{N=1}^{\infty} \bigcap_{N=n}^{\infty} A_n^c\right).$$

   If now $C_N = \bigcap_{N=n}^{\infty} A_n^c$, then $C_1 \subseteq C_2 \subseteq \ldots$ is an increasing sequence of sets.

   Further,

$$\Pr(C_N) = \lim_{m\to\infty} \Pr\left(\bigcap_{n=N}^{m} A_n^c\right) = \lim_{m\to\infty} \prod_{n=N}^{m} \left(1 - \Pr(A_n)\right),$$

   where we have used independence in forming the product. But now, since $1 - x \leq \exp(-x)$ for all $x \geq 0$, we see that

$$0 \leq \Pr(C_N) \leq \lim_{m\to\infty} \prod_{n=N}^{m} \exp(-\Pr(A_n)) = \lim_{m\to\infty} \exp\left(-\sum_{n=N}^{m} \Pr(A_n)\right) = 0,$$

   because $\sum_{n=1}^{\infty} \Pr(A_n) = \infty$.

   Hence then by continuity applied to the increasing sequence (C_N),

$$\Pr\left(\{A_n^c \text{ a.a.}\}\right) = \Pr\left(\bigcup_{N=1}^{\infty} C_N\right) = \lim_{N\to\infty} \Pr(C_N) = 0,$$

   So that $\Pr(\{A_n \text{ i.o.}\}) = 1$.

Convergence almost surely (for interest: non-examinable)

Proposition 4.18. *Let $(\Omega, \mathcal{F}, \Pr)$ be a probability space, and suppose $X_1, X_2, \ldots$ and X are random variables. We can show that the set*

$$\{X_n \to X\} = \{\omega \in \Omega : X_n(\omega) \to X(\omega)\} \in \mathcal{F},$$

i.e. $\{X_n \to X\}$ is an event.

Proof. *Recall the definition of convergence for real numbers: $x_n \to x$ means that for all $\epsilon > 0$, there exists $N(\epsilon) \geq 1$ such that whenever $n \geq N(\epsilon)$, $|x_n - x| < \epsilon$. Equivalently, for all $m \geq 1$ there exists $N(m) \geq 1$ such that whenever $n \geq N(m)$, $|x_n - x| < \frac{1}{m}$. This formulation is preferable here, as it allows us to work with countable unions and intersections.*

So then $\omega \in \{X_n \to X\}$ if and only if for all $m \geq 1$ there exists an $N(m)$ such that

$$\omega \in \bigcap_{n=N(m)}^{\infty} \left\{ |X_n(\omega) - X(\omega)| < \frac{1}{m} \right\}.$$

Translating the quantifiers into statements about subsets gives

$$\{X_n \to X\} = \bigcap_{m=1}^{\infty} \bigcup_{N(m)=1}^{\infty} \bigcap_{n=N(m)}^{\infty} \left\{ |X_n(\omega) - X(\omega)| < \frac{1}{m} \right\}.$$

By construction, this is an event.

Definition 4.19. *We say that X_n* **converges to** *X* **almost surely** *(or with probability 1), written $X_n \xrightarrow{a.s.} X$, if*

$$\Pr(X_n \to X) = 1.$$

Remark 4.20. *Almost sure convergence can be quite a subtle concept, and it is included here only to give a sense of more advanced ideas. Convergence in probability and convergence in distribution will be our focus in this module.*

Proposition 4.21. *If $X_n \xrightarrow{a.s.} X$ then $X_n \xrightarrow{P} X$.*

Proof. *By the definition of convergence, if $X_n(\omega) \to X(\omega)$ then for all $\epsilon > 0$, there exists $N \in \mathbf{N}$ such that $|X_n(\omega) - X(\omega)| < \epsilon$ for any $n \geq N$. So if $X_n \xrightarrow{a.s.} X$, then for any $\epsilon > 0$, we define the increasing sequence of events $(A_N(\epsilon))$ by*

$$A_N(\epsilon) = \{\omega \in \Omega : |X_n(\omega) - X(\omega)| < \epsilon \text{ for all } n \geq N\}.$$

Hence by the definition of convergence, $\{X_n \to X\} \subseteq \bigcup_{N=1}^{\infty} A_N(\epsilon)$, so

$$1 = \Pr(X_n \to X) \leq \Pr\left(\bigcup_{N=1}^{\infty} A_N(\epsilon) \right) \leq 1.$$

Hence, by continuity applied to the increasing sequence $(A_N(\epsilon))$,

$$\lim_{N \to \infty} \Pr(A_N(\epsilon)) = 1.$$

But now $A_N(\epsilon) \subseteq \{|X_N - X| < \epsilon\}$, so

$$\lim_{n \to \infty} \Pr(|X_N - X| < \epsilon) = 1,$$

which says that $X_N \xrightarrow{P} X$.

The strong law of large numbers (for interest: non-examinable)

Proposition 4.22. *Let $(X_n)_{n\geq 1}$ be a sequence of independent and identically distributed random variables such that $\mathrm{E}(X_i^4) < \infty$ and $\mathrm{E}(X_i) = \mu$. Then*

$$\Pr\left(\bar{X}_n \to \mu\right) = 1.$$

Proof *First, we show that there exists a constant $C > 0$ such that*

$$\mathrm{E}\left((\bar{X} - \mu)^4\right) \leq \frac{c}{n^2}.$$

To see this, define $Z_i = X_i - \mu$, $S_n = \sum_{i=1}^n X_i$, and consider

$$\mathrm{E}\left((S_n - n\mu)^4\right) = \mathrm{E}\left(\left(\sum_{i=1}^n Z_i\right)^4\right) = n\mathrm{E}(Z_1^4) + 3n(n-1)\mathrm{E}(Z_1^2 Z_2^2),$$

since all other terms in the expansion are zero, e.g.

$$\mathrm{E}(Z_1 Z_2^3) = \mathrm{E}(Z_1)\mathrm{E}(Z_2^2) = 0.$$

Now we can choose e.g. $C = 4\max\left\{\mathrm{E}(Z_1^4), \mathrm{E}(Z_1^2)^2\right\}$. Then

$$n\mathrm{E}(Z_1^4) + 3n(n-1)\mathrm{E}(Z_1^2 Z_2^2) = n^2\left(\frac{C}{4n} + \frac{3C(n-1)}{4n}\right) \leq Cn^2,$$

so that

$$\mathrm{E}\left((\bar{X} - \mu)^4\right) = \frac{1}{n^4}\mathrm{E}\left((S_n - n\mu)^4\right) \leq \frac{C}{n^2}.$$

Now we use the first Borel-Cantelli lemma to deduce the strong law of large numbers. For $\gamma > 0$, define the event

$$A_n = \left\{|\bar{X}_n - \mu| \geq n^{-\gamma}\right\}.$$

Then by Markov's inequality,

$$\Pr(|\bar{X}_n - \mu| \geq n^{-\gamma}) \leq \frac{\mathrm{E}(\bar{X}_n - \mu)^4}{n^{-4\gamma}} \leq Cn^{4\gamma-2}.$$

Note that for $\gamma < \frac{1}{4}$ we have that $\sum_{n=1}^{\infty}\Pr(A_n)$ converges, so that by the first Borel-Cantelli lemma, we get that

$$\Pr\left(\{A_n \text{ i.o.}\}\right) = \Pr\left(\bigcap_{m=1}^{\infty}\bigcup_{n\geq m} A_n\right) = 0.$$

So then we see

$$\Pr\left(\{A_n^c \text{ a.a.}\}\right) = \Pr\left(\bigcup_{m=1}^{\infty}\bigcap_{n\geq m}\left\{|\bar{X}_n - \mu| < n^{-\gamma}\right\}\right) = 1.$$

We now simply note that if $\omega \in \{A_n^c \text{ a.a.}\}$ *then there exists* $m \geq 1$ *such that for all* $n \geq m$, $\omega \in \{|\bar{X}_n - \mu| < n^{-\gamma}\}$. *Hence* $\{A_n^c \text{ a.a.}\} \subseteq \{\bar{X}_n \to \mu\}$, *and so we see that*

$$\Pr\left(\{\bar{X}_n \to \mu\}\right) = 1.$$

This is the strong law of large numbers. The conditions can be weakened somewhat, but the argument becomes rather more involved.

5
Central Limit Theorem

The central limit theorem is a result that holds in very wide generality, and this generality underpins its importance in applications. It can be thought of as a sharpening of the weak law of large numbers. As we have seen, the probability distribution of the sample mean $\bar{X}_n$ of a random sample becomes concentrated around the constant value μ, in the limit as $n \to \infty$. This says that, from the point of view of probability, $\bar{X}_n$ becomes quite uninteresting - its distribution is essentially a point mass, apart from small fluctuations. The central limit theorem 'zooms in', magnifying the fluctuations around the limit by a factor of $\sqrt{n}$ so that they become visible as a probability distribution with a well-defined density function. Remarkably, the scaled limiting distribution of the fluctuations around μ is essentially the same, regardless of the shape of the distribution of the random variables from which the original sample was taken. Whenever this parent distribution has a finite variance σ^2, we will see that the rescaled fluctuations $\sqrt{n}\,(\bar{X}_n - \mu)$ are normally distributed with variance σ^2. This approach to normality can be seen in Figure 5.1. The figure shows the distribution of the sample mean $\bar{X}_n$ for four different parent distributions, and sample size $n \in \{1, 2, 5, 25\}$.

Our approach to the central limit theorem will use moment generating functions, which we will review below. More general proofs can be given that use characteristic functions. See e.g. Billingsley (ch 27).

Moment Generating Functions

Definition 5.1. *The* **moment generating function** *of a random variable X is*

$$M_X(t) = \mathrm{E}(\exp(tX)).$$

Note that this expectation is not necessarily finite.

We note the following results for MGFs.

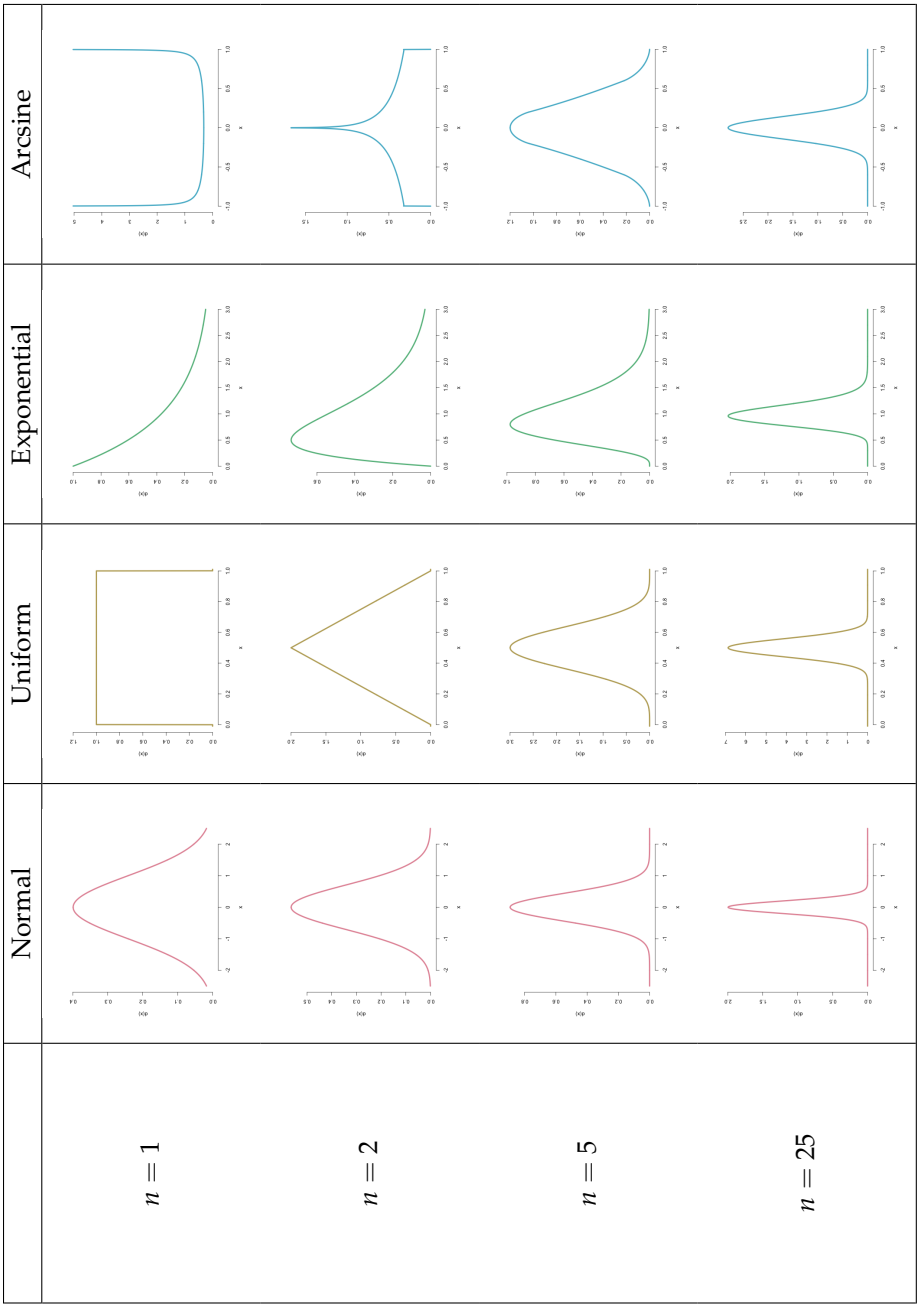


Figure 5.1: Sampling distributions of $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ where the X_i are independent and identically distributed random variables drawn from the normal, uniform, exponential and arcsine distributions. Note the convergence, in each case, of the sampling distribution to a normal distribution.

Proposition 5.2. *If $Y = aX + b$ then $M_Y(t) = \exp(bt)M_X(at)$.*
(Exercise.)

Proposition 5.3. *If X and Y are independent random variables and $Z = X + Y$, then*

$$M_Z(t) = M_X(t)M_Y(t).$$

Proof.

$$M_Z(t) = \mathrm{E}(\exp(t(X+Y)) = \mathrm{E}(\exp(tX)\exp(tY)).$$

Now, since X and Y are independent, and exp *is a continuous (and therefore measurable) function,* $\exp(tX)$ *and* $\exp(tY)$ *are independent too. It then follows that*

$$M_Z(t) = \mathrm{E}(\exp(tX))\mathrm{E}(\exp(tY)).$$

Proposition 5.4. *Suppose there exists $t_0 > 0$ such that $M_X(t) < \infty$ whenever $|t| < t_0$. Then*

$$M_X(t) = \sum_{k=0}^{\infty} \mathrm{E}(X^k)\frac{t^k}{k!},$$

and for any $k \geq 0$,

$$\frac{d^k}{dt^k}M_X(t)\mid_{t=0} = \mathrm{E}(X^k).$$

Proof. *We expand the exponential function as a power series*

$$M_X(t) = \mathrm{E}(\exp(tX)). = \mathrm{E}\left(\sum_{n=0}^{\infty}\frac{(tX)^n}{n!}\right) = \sum_{n=0}^{\infty}\mathrm{E}\left(X^n\right)\frac{t^n}{n!},$$

assuming that we may interchange the infinite summation with integration (which we can - see Billingsley, section 21). Then differentiate k times and evaluate at $t = 0$.

We note without proof the following results from analysis.

Proposition 5.5. Uniqueness *Suppose X and Y are random variables with common moment generating function $M(t)$, which is finite for $|t| < t_0$ for some $t_0 > 0$. Then X and Y are identically distributed.*

Continuity *Suppose X is a random variable with moment generating function $M_X(t)$, and $(X_n)_{n\geq 1}$ is a sequence of random variables, with respective moment generating functions $M_{X_i}(t)$. If*

$$M_{X_i}(t) \to M_X(t) < \infty$$

as $n \to \infty$ for all $|t| \leq t_0$ for some $t_0 > 0$, then $X_n \xrightarrow{D} X$.

Example 5.6. *Suppose $X \sim \Gamma(\alpha, \lambda)$, where for ease of notation λ is the rate parameter. Then we obtain the MGF of X to be*

$$\begin{aligned} M_X(t) &= \int_0^\infty \exp(tx) \frac{\lambda^\alpha x^{\alpha-1}}{\Gamma(\alpha)} \exp(-\lambda x)\, dx \\ &= \int_0^\infty \frac{\lambda^\alpha x^{\alpha-1}}{\Gamma(\alpha)} \exp(-(\lambda - t)x)\, dx \\ &= \frac{\lambda^\alpha}{(\lambda-t)^\alpha} \int_0^\infty \frac{(\lambda-t)^\alpha x^{\alpha-1}}{\Gamma(\alpha)} \exp(-(\lambda-t)x)\, dx = \left(\frac{\lambda}{\lambda - t}\right)^\alpha \qquad t < \lambda. \end{aligned}$$

Note the use of the standard trick - rewriting the integrand to take the form of a PDF, so that the integral evaluates to 1.

Remark 5.7. *In the particular case where $\alpha = 1$, we obtain the MGF for $Y \sim \text{Exp}(\lambda)$ as*

$$M_Y(t) = \frac{\lambda}{\lambda - t} \qquad t < \lambda.$$

More generally, if $\alpha = n$ is a positive integer and $Y_1, Y_2, \ldots Y_n \sim \text{Exp}(\lambda)$ are independent, then $Y = \sum_{i=1}^n Y_i$ has MGF

$$M_Y(t) = \prod_{i=1}^n \frac{\lambda}{\lambda - t} = \left(\frac{\lambda}{\lambda - t}\right)^n \qquad t < \lambda,$$

so by the uniqueness theorem, $Y \sim \Gamma(n, \lambda)$.

Example 5.8. *Let $Z \sim N(0,1)$. Then we determine $M_Z(t)$ as*

$$\begin{aligned} M_Z(t) &= \int_{-\infty}^\infty \exp(tz) \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}z^2)\, dz \\ &= \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}z^2 + tz)\, dz \end{aligned}$$

Now complete the square in the exponent

$$-\frac{1}{2}z^2 + tz = -\frac{1}{2}\left(z^2 - 2tz\right) = -\frac{1}{2}(z-t)^2 + \frac{1}{2}t^2,$$

and write the integrand so that we can identify a PDF.

$$\begin{aligned} M_Z(t) &= \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(z-t)^2 + \frac{1}{2}t^2\right) dz \\ &= \exp\left(\frac{1}{2}t^2\right) \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(z-t)^2\right) dz = \exp\left(\frac{1}{2}t^2\right). \end{aligned}$$

Remark 5.9. *The moment generating function is only really useful in the setting where the tails of the distribution decay at least as fast as an exponential. In this case, the conditions of the result above are satisfied, and we can use the MGF to determine the moments. More generally, we can use the* **characteristic function** *$\phi_X(t) = \text{E}(\exp(itX))$, which always exists.*

Remark 5.10. *An example of a heavy-tailed distribution where $M(t)$ does not exist is the* **Cauchy distribution** *with probability density function*

$$f_X(x) = \frac{1}{\pi(1+x^2)} \qquad x \in \mathbf{R}.$$

The Cauchy distribution does not have moments of any order. Recall that we define the expectation only where $\mathrm{E}|X| < \infty$. This condition is not satisfied here.

Behaviour of sample means of independent, identically distributed random variables

As a warm-up, we obtain another derivation of the weak law of large numbers in the case where the MGF is finite in some interval around zero. First, recall a useful definition for limiting behaviour of functions.

Definition 5.11. *We say $f(x) = o(g(x))$ in the limit as $x \to \infty$ if*

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = 0;$$

a similar definition is also used in the $x \to 0$ limit.

Proposition 5.12. *Suppose $X_1, X_2, \ldots$ is a sequence of independent and identically distributed random variables with common moment generating function $M(t)$, which exists in some open interval containing zero. If μ is the common expectation of the random variables, we will show that $\bar{X}_n \xrightarrow{P} \mu$.*

Proof. *If $M_n(t)$ is the MGF of $\bar{X}_n$ and $M(t)$ is the common MGF of the X_i, then by independence of the X_i,*

$$M_n(t) = \mathrm{E}\left(\exp\left(\frac{t\sum_{i=1}^n X_i}{n}\right)\right) = \prod_{i=1}^n \mathrm{E}\left(\exp\left(\frac{tX_i}{n}\right)\right) = M\left(\frac{t}{n}\right)^n.$$

Now, under the assumption that the MGF is finite in some interval around the origin, by Taylor's theorem, as $t \to 0$, we can write

$$M(t) = 1 + \mu t + o\left(t\right),$$

so that

$$M_n(t) = \left(1 + \frac{t\mu}{n} + o\left(\frac{t}{n}\right)\right)^n \to \exp(\mu t)$$

as $n \to \infty$.
Note that $\exp(\mu t)$ is the moment generating function for the constant random variable X with $\Pr(X = \mu) = 1$. Hence by continuity of moment generating functions, $\bar{X}_n \xrightarrow{D} \mu$. When the limit random variable is simply a constant, convergence in distribution and convergence in probability are equivalent, as shown in 4.14. Hence $\bar{X}_n \xrightarrow{P} \mu$.

Remark 5.13. *For a more detailed demonstration, which will also be useful in the proof of the central theorem, note that if*

$$\phi(t) = \left(1 + \frac{\alpha t}{n} + o\left(\frac{t}{n}\right)\right)^n$$

then consider

$$\log \phi(t) = n \log\left(1 + \frac{\alpha t}{n} + o\left(\frac{t}{n}\right)\right).$$

For $x \leq \frac{1}{2}$ we have $|\log(1+x) - x| \leq x^2$, so that as $n \to \infty$

$$\log \phi(t) = n\left(\frac{\alpha t}{n} + o\left(\frac{t}{n}\right)\right) \to \alpha t.$$

Then $\phi(t) \to \exp(\alpha t)$.

The central limit theorem

Proposition 5.14. *Suppose $X_1, X_2, \ldots$ is a sequence of independent, identically distributed random variables with common moment generating function $M(t)$, which exists in some open interval containing zero. Let μ and σ^2 be the common mean and variance of the X_i, respectively. Then*

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{D} Z \sim N(0,1).$$

Proof. *Let $M(t)$ be the common MGF of the centred random variables $X_i - \mu$, and $M_n(t)$ be the MGF of the standardized sample mean $Z_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$.*
Then

$$\begin{aligned} M_n(t) &= \mathrm{E}(\exp(tZ_n)) \\ &= \mathrm{E}\left(\exp\left(\frac{(t\sqrt{n}\sum_{i=1}^n (X_i - \mu)}{\sigma n}\right)\right) \\ &= \prod_{i=1}^n \mathrm{E}\left(\exp\left(\frac{t(X_i - \mu)}{\sigma\sqrt{n}}\right)\right) = M\left(\frac{t}{\sigma\sqrt{n}}\right)^n. \end{aligned}$$

In the limit $t \to 0$, we expand $M(t)$

$$\begin{aligned} M(t) &= M(0) + tM'(0) + \frac{M''(0)t^2}{2} + o\left(t^2\right) \\ &= M(0) + t\mathrm{E}(X_i - \mu) + \frac{\mathrm{E}\left[(X_i - \mu)^2\right] t^2}{2} + o\left(t^2\right) \\ &= 1 + \frac{\sigma^2 t^2}{2} + o\left(t^2\right), \end{aligned}$$

since $\mathrm{E}(X_i - \mu) = 0$ and $\mathrm{E}((X_i - \mu)^2) = \sigma^2$.

This then gives

$$M_n(t) = \left(1 + \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n,$$

so that $M_n(t) \to \exp(\frac{t^2}{2})$ *as* $n \to \infty$.

This last expression is the MGF of a standard normal variable, so that by continuity, the standardized sample mean satisfies

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{D} Z \sim N(0,1).$$

6
Stochastic Processes

Motivation

Many of the problems we have considered in probability involve sequences of independent trials. In such cases, the outcomes of earlier trials do not influence our beliefs about the outcome of later trials. There are many interesting and practically important problems that can be described as a sequence of random trials, but for which the independence assumption is not reasonable: knowing something about earlier trials changes what we believe is likely to happen next.

As a first example, we consider a simple model for the inheritance of DNA. In living organisms, genetic information is encoded in DNA as long sequences of nucleotide bases A, G, C and T. In the vast majority of sequence positions, all individuals of a species have the same base. At some sequence positions, variant bases may be introduced by mutation when the DNA is copied. If we follow a small piece of DNA as it is passed through several generations, we might see a pattern such as that below.

generation 0 $AGTTCTGTATC$
generation 1 $AGTTCTGTATC$
generation 2 $AGTTCTGTATC$
generation 3 $AGTTCTGTATC$
generation 4 $AGTTCTGTATC$
generation 5 $AGTTCAGTATC$
generation 6 $AGTTCAGTATC$

Note that the central position has experienced a mutation. Let $X_n \in \{A, G, C, T\}$ be the base in the central position in generation $n \geq 0$. Since mutations are typically rare, it will almost always be true that

$X_n = X_{n-1}$. An independent trials model would not respect this important feature of the data.

What would a simple model look like? The important property to capture is that with high probability, X_n is a copy of X_{n-1}, but just occasionally a copying error occurs. We can express this by specifying the conditional probability distribution of X_n given X_{n-1} in terms of the probability $\alpha \ll 1$:

$$\Pr(X_n = j | X_{n-1} = i) = \begin{cases} 1-\alpha & i = j \\ \frac{\alpha}{3} & \text{otherwise.} \end{cases}$$

This says that with probability $1-\alpha$, no mutation occurs, so that X_n is the same as X_{n-1}, but with probability α a mutation occurs, after which each of the three other bases is equally likely.

We can represent this model graphically as in Figure 6.1, in which the nodes represents possible states of the process and edges represent transitions between states. We think of the process as a particle that jumps randomly from state to state, according to the probabilities associated to each edge. Note the edges leaving any given node have probabilities that sum to 1.

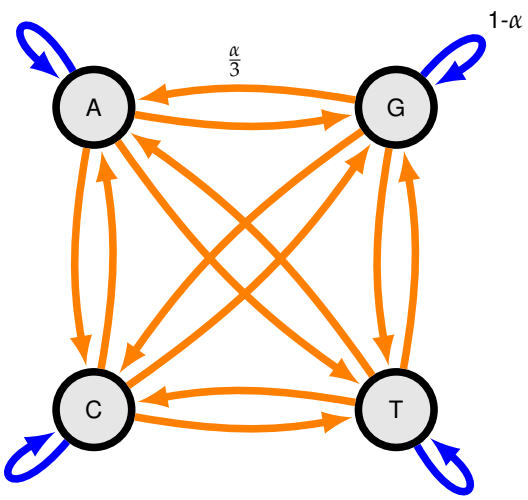


Figure 6.1: A simple model for inheritance and mutation. Orange arrows correspond to events with probability $\frac{\alpha}{3}$ and blue arrows correspond to events with probability $1-\alpha$. Note that the total flow of probability out of each node is 1.

What does this model say about the long-term behaviour of the process? Intuitively, we might think that as time passes, mutations are bound to occur, so that if we allow the process enough time, the four bases should all be equally likely. It seems that the process should eventually *forget* its initial state. How can we make this precise?

Suppose the base in the initial generation $X_0 = T$. We're interested in

$$p_n = \Pr(X_n = T | X_0 = T).$$

Certainly $p_0 = 1$, and $p_1 = 1 - \alpha$. For $n \geq 1$, we can condition on whether or not $X_{n-1} = T$:

$$\begin{aligned} p_n &= \Pr(X_n = T, X_{n-1} = T | X_0 = T) + \Pr(X_n = T, X_{n-1} \neq T | X_0 = T) \\ &= \Pr(X_n = T | X_{n-1} = T, X_0 = T) \Pr(X_{n-1} = T | X_0 = T) + \Pr(X_n = T, | X_{n-1} \neq T, X_0 = T) \Pr(X_{n-1} \neq T | X_0 = T) \\ &= \Pr(X_n = T | X_{n-1} = T, X_0 = T) p_{n-1} + \Pr(X_n = T, | X_{n-1} \neq T, X_0 = T)(1 - p_{n-1}) \end{aligned}$$

So we can write p_n in terms of p_{n-1}. It remains to find expressions for the probabilities multiplying p_{n-1} and $1 - p_{n-1}$.

To do this, we employ a key assumption about the process. The state of the process at generation n depends *only* on whether or not a mutation has occured in generation $n - 1$. For this process, it is clear that $Pr(X_n = T | X_{n-1} = T, X_0 = T) = Pr(X_n = T | X_{n-1} = T)$: if we know its state in generation $n - 1$, then knowing the state in generation 0 tells us nothing more about generation n.

We can now use the probabilities of no mutation, or of a mutation to T, to simplify the expression for p_n to

$$p_n = (1 - \alpha) p_{n-1} + \frac{\alpha}{3}(1 - p_{n-1}) = (1 - \frac{4\alpha}{3}) p_{n-1} + \frac{\alpha}{3}.$$

This now has the form of a difference equation, which we could solve by standard methods. In this simple case, we can determine the solution explicitly.

Our initial intuition was that all bases should eventually become equally likely, so that we would expect $p_n \to \frac{1}{4}$ as $n \to \infty$. Can we see this? Suppose that p_n has a limit p (which remains to be shown). What is p? Substituting the limit value in the difference equation, we must have

$$p = (1 - \frac{4\alpha}{3}) p + \frac{\alpha}{3},$$

so that indeed if $p_n \to p$, we must have $p = \frac{1}{4}$. Of course this does not establish the limiting behaviour of p_n, but it does give us a clue about where to look. Let's consider the difference between p_n and its supposed limit. Using the difference equation, this is

$$p_n - \frac{1}{4} = (1 - \frac{4\alpha}{3}) p_{n-1} + \frac{\alpha}{3} - \frac{1}{4} = \left(p_{n-1} - \frac{1}{4} \right) \left(1 - \frac{4\alpha}{3} \right).$$

This says that the difference between p_n and $\frac{1}{4}$ decreases by a factor of $1 - \frac{4\alpha}{3}$ in each generation. (What happens if $\alpha \geq \frac{3}{4}$?) Moreover, it

gives us an explicit form for p_n, for

$$p_n - \frac{1}{4} = \left(p_{n-1} - \frac{1}{4}\right)\left(1 - \frac{4\alpha}{3}\right) = \left(p_{n-2} - \frac{1}{4}\right)\left(1 - \frac{4\alpha}{3}\right)^2 = \ldots = \left(1 - \frac{1}{4}\right)\left(1 - \frac{4\alpha}{3}\right)^n,$$

where we have used the fact that $p_0 = 1$. Hence

$$p_n = \frac{1}{4} + \frac{3}{4}\left(1 - \frac{4\alpha}{3}\right)^n,$$

from which we see $p_n \to \frac{1}{4}$ as $n \to \infty$. Note that the rate at which the limit is approached is governed by the quantity $1 - \frac{4\alpha}{3}$, which is close to 1 if mutation is assumed to be very infrequent.

We have seen how, in this fairly simple model, the long-term dynamics can be determined explicitly. In what follows, we will consider models for a variety of random processes structured in time. The key property that these models will share with the example above is their *lack of memory*. We will seek answers to the following questions:

- Does the tendency of the process to forget its initial starting point hold more generally, and if so, how quickly does this forgetting occur?
- How to determine the long-term behaviour of more general models?
- What is the long-run proportion of the time spent in each state?
- How long does the process take to reach state a particular state?

Introduction

In this section, we will consider random processes on a state space $\mathcal{E}$, which will be a finite or countably infinite set. A random process is a sequence of $\mathcal{E}$-valued random variables $X_0, X_1, X_2, \ldots$. The process can be thought of as the sequence of states of a particle at the discrete times $0, 1, 2, \ldots$.

The processes we consider will all be **Markov chains**. The defining property of a Markov chain is that the particle has no memory of where it has been. The probability distribution of X_{n+1}, the particle's next state, depends on the value of X_n, its current state, and *only* on its current state, in the sense that conditioning on its more distant history does not change the distribution of the next step.

Time homogeneous Markov chains

Definition 6.1. *A* **stochastic process** *on the state space $\mathcal{E}$ is a collection of $\mathcal{E}$-valued random variables $(X_t)_{t\in\mathcal{T}}$ indexed by a set $\mathcal{T}$.*

The index set $\mathcal{T}$ should be thought of as a marker of time. We work exclusively with the discrete time case, in which $\mathcal{T} = \mathbf{N}_0 = \mathbf{N} \cup \{0\} = \{0, 1, \ldots\}$.

Definition 6.2. *The discrete time stochastic process $(X_n)_{n\in\mathbf{N}_0}$ on $\mathcal{E}$ is said to be a* **Markov chain** *if*

$$\Pr(X_n = x_n | X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) = \Pr(X_n = x_n | X_{n-1} = x_{n-1}),$$

for all $n \in \mathbf{N}$ and all $x_n, x_{n-1}, \ldots, x_0 \in \mathcal{E}$.

We can summarize the defining property of a Markov chain as saying that the future is conditionally independent of the past, given the present.

For many systems of practical interest, the Markov property arises as a commonly used modelling assumption.

- In finance, it is related to the **efficient markets hypothesis**. This says that the current price of an asset reflects all information known to the market.
- In physics, the future motion of a particle is determined by its current state (its position and velocity), through the laws of motion. The Markov property is a natural generalization of this assumption: we assume that the **probability distribution** of a particle's future motions is determined by its current state.

Definition 6.3. *A Markov chain is* **time homogeneous** *if*

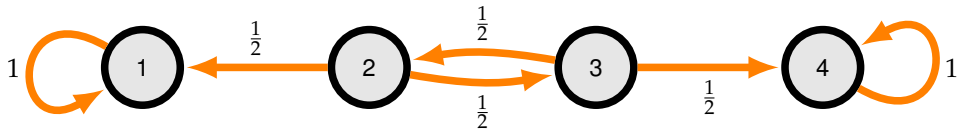
$$\Pr(X_{n+1} = j | X_n = i) = \Pr(X_1 = j | X_0 = i)$$

for all $n \in \mathbf{N}_0$ *and all* $i, j \in E$.

All of the Markov chains we work with will be time homogeneous.

Definition 6.4. *The matrix* $P = (p_{ij})_{i,j \in \mathcal{E}}$, *where* $p_{ij} = \Pr(X_1 = j | X_0 = i)$ *is the* **transition matrix** *for the time homogeneous Markov chain* (X_n).

Note that the transition matrix of a Markov chain is a **stochastic matrix**: all entries are non-negative, and the entries of each row sum to 1.



Example 6.5. *The simple random walk on* $\mathcal{E} = \{1, 2, 3, 4\}$ *with absorbing boundaries at both ends. The transition matrix for this chain is*

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

The initial distribution

The transition matrix specifies how the chain moves from one state to another. To specify the stochastic process fully, we must also specify its **initial distribution** $\lambda = (\lambda_j)_{j \in \mathcal{E}}$, where $\lambda_j = \Pr(X_0 = j)$.

Once the initial distribution and transition matrix of the chain have been specified, the distribution of the process is specified for all future times. For example, the marginal distribution of X_1 is determined by the initial distribution λ and P. Using the law of total probability,

$$\Pr(X_1 = j) = \sum_{i \in E} \Pr(X_1 = j | X_0 = i) \Pr(X_0 = i) = \sum_{i \in E} p_{ij} \lambda_i.$$

Similarly, the joint distribution of (X_2, X_1) is given by

$$\Pr(X_2 = k, X_1 = j) = \Pr(X_2 = k | X_1 = j) \Pr(X_1 = j) = p_{jk} \sum_{i \in E} p_{ij} \lambda_i.$$

Independence versus conditional independence

The Markov assumption relates to **conditional independence**. It is true to say that that if (X_n) is a Markov chain then X_n is conditionally independent of X_{n-2}, given X_{n-1}. This means that for all $x, y, z \in \mathcal{E}$

$$\Pr(X_n = z, X_{n-2} = x | X_{n-1} = y) = \Pr(X_n = z | X_{n-1} = y) \Pr(X_{n-2} = x | X_{n-1} = y).$$

However, in general X_2 and X_0 are not independent. We can see this by considering the absorbing random walk above.

Suppose the initial distribution of the chain is $\lambda = \left(\frac{1}{2}, 0, 0, \frac{1}{2}\right)$. We will show that X_2 and X_0 are not independent, by showing that

$$\Pr(X_2 = 4, X_0 = 1) = 0 \neq \Pr(X_2 = 4) \Pr(X_0 = 1).$$

It is clear that $\Pr(X_2 = 4, X_0 = 1) = 0$, because the probability of ever leaving the state 1 is zero. To see this formally,

$$\begin{aligned} \Pr(X_2 = 4, X_0 = 1) &= \Pr(X_2 = 4 | X_0 = 1) \Pr(X_0 = 1) \\ &= \sum_{j \in \mathcal{E}} \Pr(X_2 = 4, X_1 = j | X_0 = 1) \Pr(X_0 = 1). \end{aligned}$$

Now consider the terms of the sum. Using the definition of conditional probability and the Markov property,

$$\begin{aligned} \Pr(X_2 = 4, X_1 = j | X_0 = 1) &= \Pr(X_2 = 4 | X_1 = j, X_0 = 1) \Pr(X_1 = j | X_0 = 1) \\ &= \Pr(X_2 = 4 | X_1 = j) \Pr(X_1 = j | X_0 = 1). \end{aligned}$$

If $j = 1$ then the first expression is zero, whereas if $j \neq 1$, the second expression is zero. Hence $\Pr(X_2 = 4, X_0 = 1) = 0$.

We now compute the marginal probabilities.

$$\Pr(X_2 = 4) = \sum_{j \in \mathcal{E}} \sum_{i \in \mathcal{E}} \Pr(X_2 = 4, X_1 = j, X_0 = i).$$

Again considering each term and applying the Markov property,

$$\begin{aligned} \Pr(X_2 = 4, X_1 = j, X_0 = i) &= \Pr(X_2 = 4 | X_1 = j, X_0 = i) \Pr(X_1 = j | X_0 = i) \Pr(X_0 = i) \\ &= \Pr(X_2 = 4 | X_1 = j) \Pr(X_1 = j | X_0 = i) \Pr(X_0 = i) \\ &= \begin{cases} \frac{1}{2} & i = j = 4 \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Hence $\Pr(X_2 = 4) = \frac{1}{2}$. We see immediately from the initial distribution that $\Pr(X_0 = 1) = \frac{1}{2}$, and so

$$\Pr(X_2 = 4, X_0 = 1) \neq \Pr(X_2 = 4) \Pr(X_0 = 1).$$

We conclude that X_0 and X_2 are not independent, even though they are conditionally independent given X_1.

Simple random walk

Example 6.6. *The simple random walk . Let $\mathcal{E} = \mathbf{Z}$ and define the Markov chain $(X_n)_{n \in \mathbf{N}_0}$ by $X_0 = 0$ and*

$$p_{ij} = \begin{cases} p & \text{if } j = i + 1 \\ 1 - p & \text{if } j = i - 1 \\ 0 & \text{otherwise.} \end{cases}$$

Realizations from the simple random walk are plotted in Figure 6.3.

How far will the chain have moved after n time points? We seek the probability $\Pr(X_n = j | X_0 = 0)$. In this case, note that X_n can be written as

$$X_n = \sum_{i=1}^{n} Z_i,$$

where the random variables Z_i are independent and identically distributed, with

$$\Pr(Z_i = k) = \begin{cases} p & \text{if } k = 1 \\ 1 - p & \text{if } k = -1 \\ 0 & \text{otherwise.} \end{cases}$$

Suppose that u of the Z_i take the value 1 and the remaining $d = n - u$ take the value -1. Note that $X_n = j$ if and only if $j = u - d$. Solving for u and d in terms of n and j gives

$$u = \frac{n + j}{2} \text{ and } d = \frac{n - j}{2}.$$

Hence we see that

$$\Pr(X_n = j | X_0 = 0) = \begin{cases} \binom{n}{\frac{n+j}{2}} p^{\frac{n+j}{2}} (1 - p)^{\frac{n-j}{2}} & n + j \text{ even} \\ 0 & \text{otherwise.} \end{cases}$$

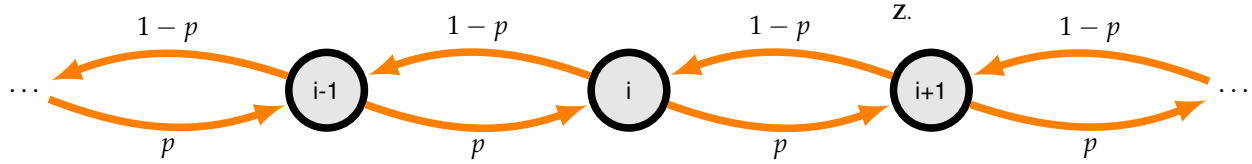


Figure 6.2: The simple random walk on **Z**.

n-step transition probabilities

Definition 6.7. *For a Markov chain $(X_n)_{n\in\mathbf{N}_0}$, the matrix $P(n)$ with entries*

$$p_{ij}(n) = \Pr(X_n = j | X_0 = i)$$

is the matrix of n-step transition probabilities.

Remark 6.8. *Clearly $P(1) = P$, the transition matrix of the chain, and $P(0) = I$, the identity matrix.*

Proposition 6.9. *The Chapman-Kolmogorov equations. Suppose $m \geq 0$ and $n \geq 1$, then*

$$p_{ij}(m+n) = \sum_{l\in E} p_{il}(m)p_{lj}(n).$$

As matrices,

$$P(m+n) = P(m)P(n),$$

from which we deduce that $P(m) = P^m$, the mth power of the transition matrix.

To prove this, we first use the law of total probability.

$$p_{ij}(m+n) = \Pr(X_{m+n} = j | X_0 = i) = \sum_{l\in E} \Pr(X_{m+n} = j, X_m = l | X_0 = i).$$

By a standard property of conditional probability, we have

$$p_{ij}(m+n) = \sum_{l\in E} \Pr(X_{m+n} = j | X_m = l, X_0 = i) \Pr(X_m = l | X_0 = i).$$

Now by the Markov property,

$$\Pr(X_{m+n} = j | X_m = l, X_0 = i) = \Pr(X_{m+n} = j | X_m = l).$$

Further, using time homogeneity we see that

$$\Pr(X_{m+n} = j | X_m = l) = \Pr(X_n = j | X_0 = l),$$

giving

$$\begin{aligned} p_{ij}(m+n) &= \sum_{l\in E} \Pr(X_n = j | X_0 = l) \Pr(X_m = l | X_0 = i) \\ &= \sum_{l\in E} p_{lj}(n)p_{il}(m), \end{aligned}$$

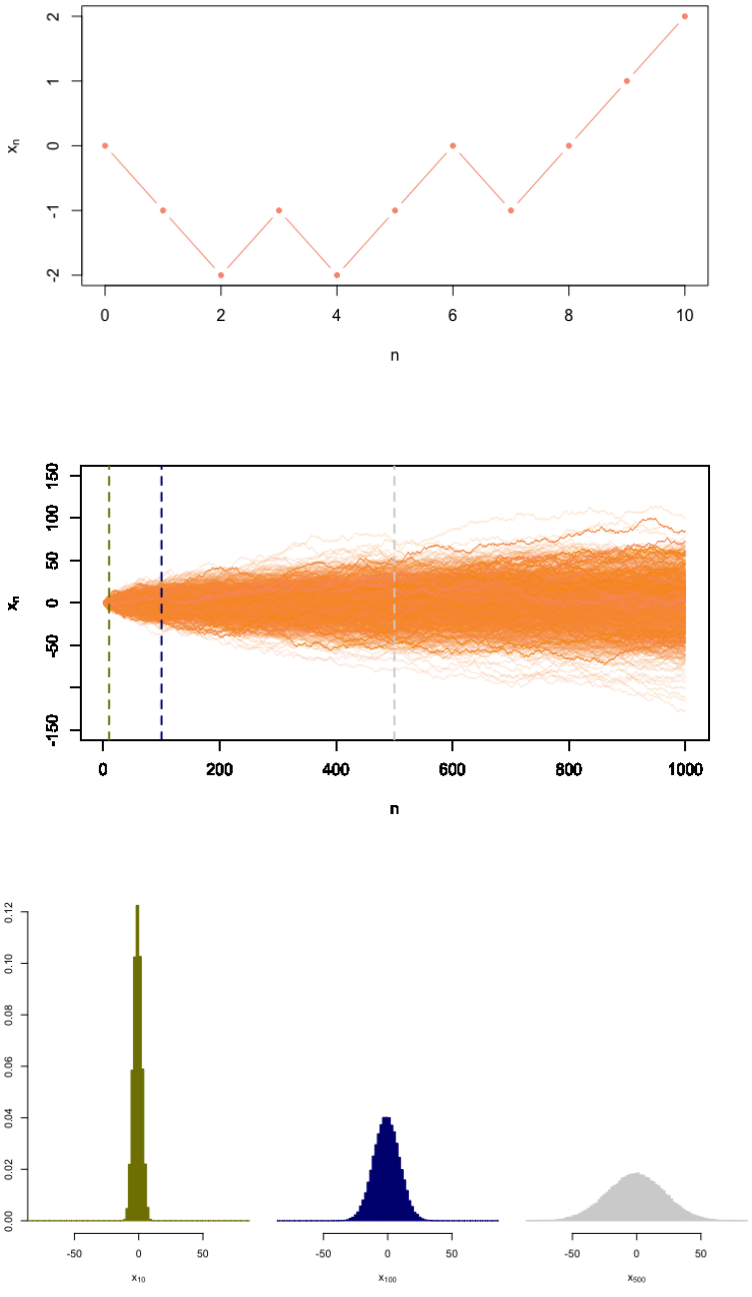


Figure 6.3: (Top) One realization of the simple random walk X_n for $n = 0$ to 10. (Centre) Many realizations of X_n started from $X_0 = 0$. (Bottom) The marginal distributions of the realizations for $n = 10, 100$ and 500.

as required.

Suppose $|\mathcal{E}| = K$ (with the obvious extension if $\mathcal{E}$ is countably infinite). Recalling the definition of matrix multiplication, we see that the (i,j)th entry of the matrix P^2 is

$$\sum_{l=1}^{K} p_{il}p_{lj} = p_{ij}(2).$$

This argument extends immediately by induction to show that the (i,j)th element of P^n is $p_{ij}(n)$ for any $n \geq 1$. Note also that $P^0 = I$, by the definition of conditional probability.

Example 6.10. *The two-state Markov chain. The transition diagram of the most general two-state Markov chain is shown in Figure 6.4. Its transition matrix is*

$$\begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}.$$

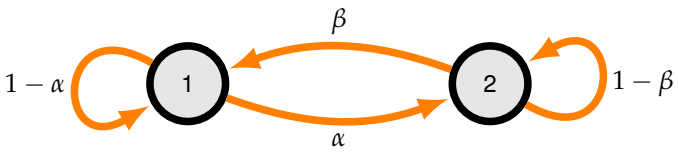


Figure 6.4: The general two-state Markov chain with $\Pr(X_1 = 1|X_0 = 2) = \beta$ and $\Pr(X_1 = 2|X_0 = 1) = \alpha$

We can compute $p_{ij}(n)$ explicitly for this model. From the matrix formulation of the Chapman-Kolmogorov equations, we know that $P^{n+1} = P^n P$. Considering individual entries,

$$p_{11}^{(n+1)} = p_{12}^{(n)}\beta + p_{11}^{(n)}(1-\alpha),$$

and since $p_{11}^{(n)} + p_{12}^{(n)} = 1$, elimination gives a recurrence relation satisfied by $p_{11}^{(n)}$:

$$p_{11}^{(n+1)} = (1 - p_{11}^{(n)})\beta + p_{11}^{(n)}(1-\alpha) = \beta + (1-\alpha-\beta)p_{11}^{(n)}.$$

This recurrence relation has a unique solution subject to the condition that $p_{11}^0 = 1$, given by

$$p_{11}(n) = \frac{\beta}{\alpha+\beta} + \frac{\alpha}{\alpha+\beta}(1-\alpha-\beta)^n.$$

Then it follows by symmetry that

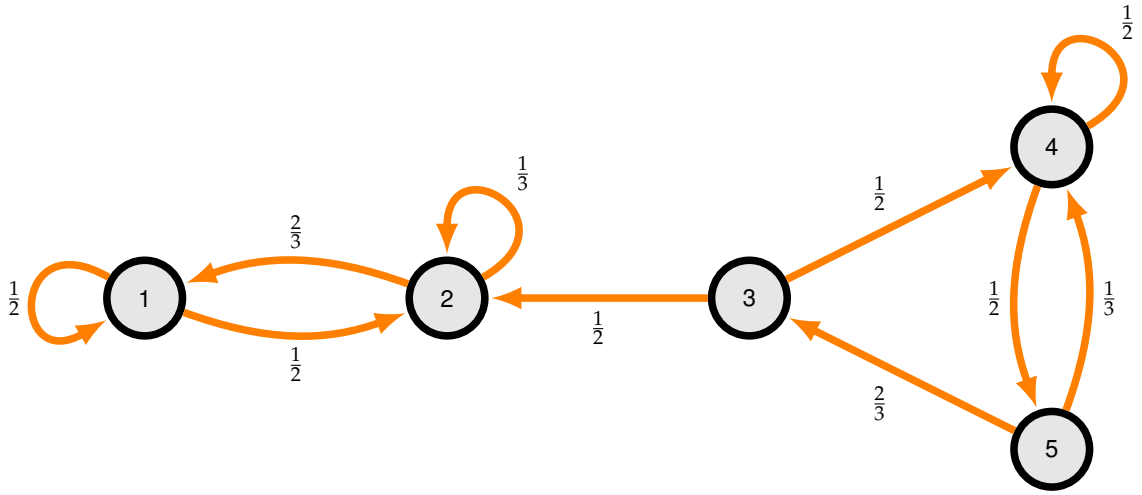
$$p_{22}(n) = \frac{\alpha}{\alpha+\beta} + \frac{\beta}{\alpha+\beta}(1-\alpha-\beta)^n,$$

and the off-diagonal entries of P^n can be computed as complementary probabilities.

An alternative approach proceeds via computation of the eigenvectors of P: see problem sheet.

Class structure

Figure 6.5: A five-state Markov chain.



As soon as we get beyond two states, the behaviour of a Markov chain can become difficult to describe explicitly. We will see in this section that the chain itself suggests a natural decomposition of the state space into components, on which its behaviour is more easily understood.

Consider the Markov chain whose transition diagram is shown in Figure 6.5. Its transition matrix is

$$\begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{2}{3} & \frac{1}{3} & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & \frac{2}{3} & \frac{1}{3} & 0 \end{pmatrix}.$$

Note that while transitions from state 3 to state 2 are possible, it is not possible to make a transition in the reverse direction. If the chain starts in one of the states $3, 4, 5$, eventually - if we wait long enough - it will end up jumping via state 3 into state 2. Once this transition has happened, the chain is stuck on the left hand side: it moves between states 1 and 2 forever.

Definition 6.11. *The state j is said to be* **accessible** *from the state i, written $i \to j$, if there exists $n \geq 0$ such that $p_{ij}(n) > 0$. So j is accessible from i when there is positive probability that, starting at i, the chain ever reaches j. Note that it is often easier to read $i \to j$ as i* **leads to** *j.*

Example 6.12. *In Figure 6.5, all states are accessible from state* 3*. Only states* 1 *and* 2 *are accessible from state* 2*.*

Definition 6.13. *The states i and j are said to* **communicate***, written $i \leftrightarrow j$ if $i \to j$ and $j \to i$.*

Example 6.14. *In Figure 6.5, state* 3 *communicates with states* 4 *and* 5*; these two states also communicate with each other. States* 1 *and* 2 *communicate.*

Proposition 6.15. *The binary relation $i \leftrightarrow j$ is an equivalence relation on $\mathcal{E}$.*

Proof. *The relation is immediately seen to be* **reflexive***, since for any $i \in \mathcal{E}$, we have that*

$$p_{ii}(0) = \Pr(X_0 = i | X_0 = 0) = 1 > 0.$$

The relation is clearly **symmetric** *from its definition in terms of $i \to j$ and $j \to i$.*

To see that the relation is **transitive***, suppose that $i, j, k \in \mathcal{E}$ are distinct states such that $i \leftrightarrow j$ and $j \leftrightarrow k$. Then there exist $m, n \geq 0$ such that $p_{ij}(m) > 0$ and $p_{jk}(n) > 0$.*

Since the states are distinct, m and n are strictly positive and by the Chapman-Kolmogorov equations,

$$p_{ik}(m+n) = \sum_{l \in \mathcal{E}} p_{il}(m) p_{lk}(n) \geq p_{ij}(m) p_{jk}(n) > 0$$

Remark 6.16. *The equivalence relation$\leftrightarrow$ partitions the sample space E into* **communicating classes** *.*

Definition 6.17. *A set of states C is* **closed** *if $p_{ij} = 0$ for all $i \in C$ and $j \notin C$.*

Remark 6.18. *Informally, a closed class is one from which the chain cannot escape.*

Definition 6.19. *A set of states C is said to be* **irreducible** *if $i \leftrightarrow j$ for all $i, j \in C$. A Markov chain the state space $\mathcal{E}$ is irreducible if its entire state space $\mathcal{E}$ is irreducible.*

Periodicity

Recall the simple random walk on the integers: a chain with state space $\mathcal{E} = \mathbf{Z}$ which jumps from i to $i+1$ with probability p and to $i-1$ with probability $1-p$. Note that this chain *must* move at each time step - it cannot stay where it is. This gives us quite a bit of information about where the chain can be at future times. For

instance, suppose we start the chain in state $X_0 = i$. At what future times is it possible for the chain to return to its starting point?
We have seen that the random walk can be written as a sum of independent, identically distributed increments

$$X_n = i + \sum_{k=1}^{n} Z_k, \qquad Z_k \in \{-1, 1\}.$$

If $X_n = i$, then the increments must sum to zero: there must be as many 1s as -1s in the sequence, so n must be an even number. The chain has zero probability of returning to its starting point at odd-numbered times. This is quite a restriction on its behaviour.

We can observe similar behaviour in a degenerate case of the two-state chain of Figure 6.4 if we take $\alpha = 1 = \beta$, so that the transition matrix is

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Again, the chain *must* jump at each step. What does this say about the $n - step$ transition probabilities?

If $n = 2k$ is even, then (as you can check),

$$P^n = \left(P^2\right)^k = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}^k = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

but if $n = 2k + 1$ is odd, then

$$P^n = P^{2k}P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

So again in this example, we see that the chain can only return to its starting point at even times (in fact in this case, it *always* returns to its starting point at even times).

In larger chains, more complex behaviour is possible. Consider the seven-state chain in Figure 6.6. If the chain starts in state 7, at what future times can it return to its starting point? What about state 4?

Definition 6.20. *The* **period** *of the state i is $d(i) = \gcd\{n > 0 : p_{ii}(n) > 0\}$, the greatest common divisor of the times at which a return to i is possible. If $d(i) = 1$, the state i is said to be* **aperiodic**, *and* **periodic** *if $d(i) > 1$.*

Example 6.21. *In Figure 6.6, states 1 to 4 have period 2. To see this, suppose that $X_0 = 4$. At the next time step, the chain is equally likely to be in states*

1 *or* 3*. By symmetry, in either case,* X_2 *is either in states* 2 *or* 4*, and these must have equal probability. So* $p_{44}(2) = \frac{1}{2}$*. Arguing by induction, we can see that in fact*

$$p_{44}(n) = \begin{cases} \frac{1}{2} & n \text{ even} \\ 0 & n \text{ odd.} \end{cases}$$

The same result can be obtained for states 1, 2 *and* 3*. The remaining states have period* 3*, e.g. for state* 7*, the only path back to itself with non-zero probability is* $7 \to 6 \to 5 \to 7$*. Hence only if* $3|n$ *can we have* $p_{77}(n) > 0$*.*

Proposition 6.22. *All states in the same communicating class have the same periodicity.*

Proof. *Suppose that* $i \leftrightarrow j$ *and* $d|n$ *whenever* $p_{ii}(n) > 0$*. We want to show that if* $p_{jj}(m) > 0$ *for some* $m > 0$*, then* $d|m$*.*

Since $i \leftrightarrow j$*, we can find* $a, b \geq 0$ *such that* $p_{ij}(a) > 0$ *and* $p_{ji}(b) > 0$*. But this then gives a path from* i *to itself in* $a + b$ *steps: by the Chapman-Kolmogorov equations,*

$$p_{ii}(a+b) = \sum_{l \in \mathcal{E}} p_{il}(a) p_{li}(b) \geq p_{ij}(a) p_{ji}(b) > 0,$$

so that we must have $d|a + b$*.*

Now since $p_{jj}(m) > 0$*, we can travel from* i *back to itself in* $a + m + b$ *steps:*

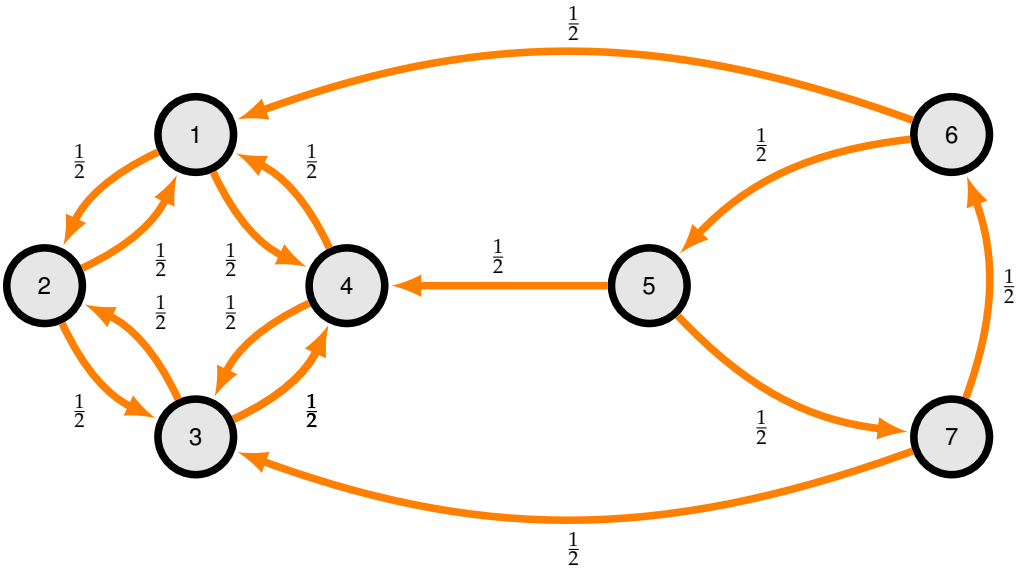
$$i \xrightarrow{a} j \xrightarrow{m} j \xrightarrow{b} i,$$

so that we must have $p_{ii}(a + m + b) > 0$*, and so we see that* $d|a + m + b$*. Since we know already that* $d|a + b$*, we must also have that* $d|m$*.*

We have shown that the sets $\{n : p_{ii}(n) > 0\}$ *and* $\{m : p_{jj}(m) > 0\}$ *share the same divisors, and so must share the same greatest common divisor.*

Remark 6.23. *Periodicity and irreducibility are said to be* **structural properties** *i of the chain. They depend only on how the chain is connected, i.e. on which transitions have strictly positive probability.*

Figure 6.6: A seven-state Markov chain.



Classification of States

Consider again the seven-state chain in Figure 6.6. The elements of the two communicating classes are different from each other in several ways. As we have already seen, the four states on the left all have period 2, whereas the three on the right have period 3.

Apart from this, there is another fundamental difference between the elements of the two classes, in terms of the chain's behviour at arbitrarily long times.

To see this, first suppose the chain starts in state 7. Whatever happens, we know it moves to a different state at the next time step. Is the chain guaranteeed to return to state 7? It is by no means certain. In fact, if any of the following transitions occur:

$$7 \to 6 \to 1, \qquad 7 \to 6 \to 5 \to 4, \qquad 7 \to 3,$$

then the chain can *never* return to its starting point. Since each of these transitions has positive probability, we see that the probability of ever returning to state 7 is strictly smaller than 1.

If we instead start the chain in state 4, the situation is rather different. The only way that the chain could avoid returning to its starting point is by alternately jumping between the set $\{1,3\}$ at odd times and the state 2 at even times. This means it is guaranteed to return to state 4 eventually, with probability 1.

Let's argue more formally. For $n \geq 1$, let A_n be the event that the chain returns to state 4 at or before time n. Note that (A_n) is an increasing sequence of events. We're interested in

$$p_n = \Pr(A_n | X_0 = 4).$$

Certainly $p_1 = 0$ and for $k \geq 0$, $p_{2k+1} = p_{2k}$, because the chain can only visit state 4 at even times.

We now determine p_{2k}, for $k \geq 1$ by considering the complementary event. If A_{2k} has not occurred, then the path of the chain starting at time 0 must be

$$4 \to \{1,3\} \to 2 \to \{1,3\} \to \ldots \to 2.$$

We can use the Markov property to determine the required probability. Note that all of the transitions $2 \to \{1,3\}$ have probability 1, and each of the k transitions from $\{1,3\}$ to 2 has probability $\frac{1}{2}$.

This then gives

$$p_{2k} = 1 - \frac{1}{2^k}.$$

Now let A be the event that the chain *ever* reaches state 4, so that

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

Then as (A_n) is an increasing sequence, we see that

$$\Pr(A | X_0 = 4) = \lim_{n \to \infty} \Pr(A_n | X_0 = 4) = 1.$$

Hence if the chain starts in state 4, it returns to its original state with probability 1. Such states are said to be *recurrent*.

Definition 6.24. *A state $i \in \mathcal{E}$ is said to be* **recurrent** *for the Markov chain X_n if the probability that the chain, having started in state i, ever returns to i, is equal to* 1*:*

$$\Pr(X_n = i, \textit{ for some } n \geq 1 | X_0 = i) = \Pr\left(\bigcup_{n=1}^{\infty} \{X_n = i\} | X_0 = i \right) = 1,$$

and **transient** *if this probability is smaller than one,*

$$\Pr\left(\bigcup_{n=1}^{\infty} \{X_n = i\} | X_0 = i \right) < 1.$$

Definition 6.25. *The* **first passage time** *of the state $j \in \mathcal{E}$ is*

$$T_j = \min\{n \geq 1 : X_n = j\},$$

the first n such that $X_n = j$.

Remark 6.26. *The first passage time is not necessarily finite: $T_j = \infty$ if the chain never visits state j. This means that it is not strictly correct to refer to T_j as a random variable - it is not necessarily real-valued. We will only really use T_j in that we consider $\{T_j = n\}$ to be shorthand for the event*

$$\{X_n = j, X_{n-1} \neq j, \ldots, X_1 \neq j\}.$$

Remark 6.27. *For states $i, j \in \mathcal{E}$, we denote the probability of the event $\{T_j = n\}$, conditional on starting in state i, as*

$$f_{ij}(n) := \Pr\left(T_j = n | X_0 = i\right) = \Pr(X_n = j, X_{n-1} \neq j, \ldots, X_1 \neq j | X_0 = i),$$

and similarly $f_{ij} = \Pr(T_j < \infty | X_0 = i)$.

Noting that the events $\{T_j = n\}$ are disjoint, we have that

$$f_{ij} = \Pr\left(\bigcup_{n=1}^{\infty} \{T_j = n\} | X_0 = i\right) = \sum_{n=1}^{\infty} f_{ij}(n).$$

This is the probability that the chain ever hits the state j given that it starts from state i.

Remark 6.28. *Note that the state i is recurrent if and only if $f_{ii} = 1$, and transient if and only if $f_{ii} < 1$. A Markov chain must in fact visit each recurrent state i infinitely often, because once it reaches i, as it must, it is just as if the process restarts, starting from i.*

What happens if i is a transient state? Once again, if the chain hits i, then it is as if the process begins anew, starting from i. But for a transient state, there is a probability $1 - f_{ii}$ that it will never return. Since the chain has no memory of its prior visits, the random variable N_i, which counts the chain's visits to the state i, has a geometric distribution with success probability $1 - f_{ii}$. In particular, N_i is finite with probability 1.

Proposition 6.29. *For states $i, j \in E$ and $n \geq 1$,*

$$p_{ij}(n) = \sum_{l=1}^{n} f_{ij}(l) p_{jj}(n-l).$$

In particular, $p_{ij} = p_{ij}(1) = f_{ij}(1)$.

Proof *Note first that the event $\{X_n = j\}$ can be decomposed into the disjoint union*

$$\{X_n = j\} = \bigcup_{l=1}^{n} \{X_n = j, T_j = l\},$$

The law of total probability then gives

$$\Pr(X_n = j | X_0 = i) = \sum_{l=1}^{n} \Pr(X_n = j,\ T_j = l | X_0 = i) = \sum_{l=1}^{n} \Pr(X_n = j | T_j = l,\ X_0 = i) \Pr(T_j = l | X_0 = i).$$

We now apply the definition of T_j and then use the Markov property:

$$\Pr(X_n = j | T_j = l,\ X_0 = i) = \Pr(X_n = j | X_l = j, X_{l-1} \neq j, \ldots, X_1 \neq j,\ X_0 = i) = \Pr(X_n = j | X_l = j).$$

This then gives

$$\Pr(X_n = j | X_0 = i) = \sum_{l=1}^{n} \Pr(X_n = j | X_l = j) \Pr(T_j = l | X_0 = i) = \sum_{l=1}^{n} p_{jj}(n-l) f_{ij}(l).$$

Proposition 6.30. *The state i is recurrent if and only if $\sum_{n=1}^{\infty} p_{ii}(n) = \infty$; equivalently, the state i is transient if and only if $\sum_{n=1}^{\infty} p_{ii}(n) < \infty$.*

Proof. *Let $I(A)$ be the indicator random variable for the event A, i.e. the random variable taking values 1 and 0 according to whether or not the event A occurs.*

Then the random variable

$$N_i = \sum_{n=1}^{\infty} I\{X_n = i\}$$

counts the number of visits of the Markov chain $(X_n)_{n \in \mathbf{N}_0}$ to the state i.

$$\mathrm{E}(N_i | X_0 = i) = \sum_{n=1}^{\infty} \mathrm{E}\left(I\{X_n = i\} | X_0 = i\right) = \sum_{n=1}^{\infty} p_{ii}(n).$$

By the earlier remark, if i is recurrent then the chain returns to i infinitely often with probability 1, so that $\mathrm{E}(N_i | X_0 = i)$ is infinite.

In contrast, if i is transient, then N_i follows a geometric distribution, so that

$$\Pr(N_i = k | X_0 = i) = f_{ii}^k (1 - f_{ii}), \qquad k \geq 0,$$

and so

$$\mathrm{E}(N_i | X_0 = i) = \sum_{k=0}^{\infty} k \Pr(N_i = k | X_0 = i) = \sum_{k=0}^{\infty} k f_{ii}^k (1 - f_{ii}) = \frac{f_{ii}}{1 - f_{ii}} < \infty.$$

Remark 6.31. *The last equality follows on using the expectation of a geometric random variable.*

Note that we condition on $X_0 = i$ in all probabilities. So we could have defined the number of visits to state i including the visit at time 0. In this case, N_i takes values $k \geq 1$, and has a shifted geometric probability distribution. Recall that there are two forms for the geometric distribution: one that counts the number of trials before a success, and the other that counts the total number of trials. The two corresponding random variables are different by 1.

Proposition 6.32. *For states $i \leftrightarrow j$, either i and j are both recurrent, or they are both transient.*

Proof. *Suppose $i \leftrightarrow j$. Then there exist $m, n \geq 0$ such that $p_{ij}(m) > 0$ and $p_{ji}(n) > 0$.*
For any $l \geq 0$, the Chapman-Kolmorogorov equations give

$$p_{jj}(m+l+n) = \sum_{k \in \mathcal{E}} p_{jk}(m+l) p_{kj}(n) \geq p_{ji}(m+l) p_{ij}(n).$$

Using the Chapman-Kolmogorov equations again, we get

$$p_{ji}(m+l) = \sum_{k \in \mathcal{E}} p_{jk}(m) p_{ki}(l) \geq p_{ji}(m) p_{ii}(l),$$

giving

$$p_{jj}(m+l+n) \geq p_{ji}(m) p_{ii}(l) p_{ij}(n).$$

So now we see that

$$\sum_{l=1}^{\infty} p_{jj}(l) \geq \sum_{l=1}^{\infty} p_{jj}(m+l+n) \geq p_{ji}(m) p_{ij}(n) \sum_{l=1}^{\infty} p_{ii}(l).$$

If now i is a recurrent state, then the sum on the right hand side diverges. Hence the sum on the left must also diverge, so that j is also recurrent. i and j are symmetrical in the argument above, so this completes the proof.

Proposition 6.33. *Let C be a recurrent communicating class. Then C is closed: for $i \in C$ and $j \notin C$, we must have $p_{ij} = 0$.*

Proof. *Suppose for contradiction that $p_{ij} > 0$.*
Since $j \notin C$, $j \nrightarrow i$, so that we have the inclusion of events

$$\{X_1 = j\} \subseteq \bigcap_{n=1}^{\infty} \{X_n \neq i\}.$$

This says that if $X_1 = j$ then the chain never reaches state i. So then conditioning on starting in state i,

$$0 < \Pr(X_1 = j | X_0 = i) \leq \Pr\left(\bigcap_{n=1}^{\infty} \{X_n \neq i\} | X_0 = i \right).$$

But this contradicts the recurrence of state i, since now

$$\Pr(X_n = i, \text{ for some } n \geq 1 | X_0 = i) = 1 - \Pr\left(\bigcap_{n=1}^{\infty} \{X_n \neq i\} | X_0 = i \right) < 1.$$

we now have a useful result as a corollary.

Proposition 6.34. *The state space decomposes uniquely as*

$$\mathcal{E} = T \cup C_1 \cup C_2 \cup \dots,$$

where T is the set of transient states, and $C_1, C_2, \dots$ are irreducible, closed sets of recurrent states.

Remark 6.35. *This last result is very useful in simplifying the dynamics of a Markov chain. For, if the chain starts in one of the sets C_i, then it stays there forever, so we may as well take this set to be the entire state space. If the chain starts in T, then either it stays in T forever, or it jumps at some point to one of the C_i, and remains in C_i for evermore. Where the state space $\mathcal{E}$ is finite, the chain can only spend a finite amount of time in the set T of transient states.*

Definition 6.36. *The* **mean recurrence time** *of the state $i \in \mathcal{E}$ is $\mu_i = \mathrm{E}(T_i | X_0 = i)$.*

Remark 6.37. *For transient states, $\Pr(T_i = \infty | X_0 = i) > 0$, so that for such states, necessarily $\mu_i = \infty$.*

For recurrent states, $\Pr(T_i < \infty | X_0 = i) = 1$, however, the expectation

$$\mathrm{E}(T_i | X_0 = i) = \sum_{n=1}^{\infty} n f_{ii}(n)$$

may be finite or infinite in general.

Definition 6.38. *The recurrent state $i \in \mathcal{E}$ is said to be* **null recurrent** *if $\mu_i = \infty$ and positive recurrent if $\mu_i < \infty$.*

Hitting times

Definition 6.39. *If (X_n) is a Markov chain on $\mathcal{E}$, then the hitting time of a set $A \subseteq \mathcal{E}$ is the random variable*

$$H^A = \min\{n \geq 0 : X_n \in A\}.$$

We take the min *of the empty set to be ∞; this corresponds to a chain that never reaches A.*
If the chain $(X_n)_{n\in\mathbf{N}_0}$ starts at $i \in \mathcal{E}$, then we define the hitting probability

$$h_i^A = \Pr(H^A < \infty | X_0 = i).$$

Remark 6.40. *We will often be interested in quantities h_i^A where $A = \{j\}$ - the probability that a chain starting in state $i \in \mathcal{E}$ reaches a state $j \in \mathcal{E}$. To simplify notation, we denote this by h_i^j.*

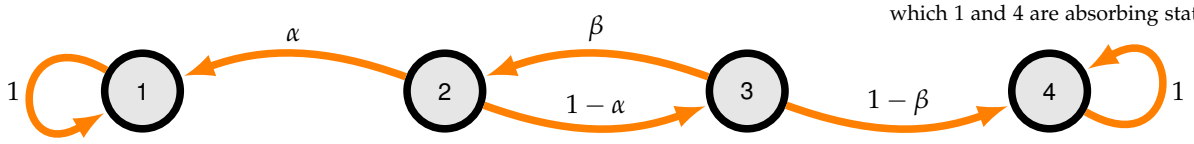


Figure 6.7: A four-state Markov chain in which 1 and 4 are absorbing states.

Example 6.41. *Consider the Markov chain with $E = \{1, 2, 3, 4\}$ shown in Figure 6.7, which has the transition matrix*

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \alpha & 0 & 1-\alpha & 0 \\ 0 & \beta & 0 & 1-\beta \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Clearly states 1 and 4 are absorbing - once the chain reaches such a state, it never leaves.

Suppose the chain starts at $X_0 = 2$. What is the probability of absorption at 4?

We consider the first step of the chain after it starts. With probability α it moves to 1, and with probability $1 - \alpha$ it moves to 3. Using the law of total probability, we can write

$$\begin{aligned} h_2^4 &= \Pr(H^4 < \infty, X_1 = 1 | X_0 = 2) + \Pr(H^4 < \infty, X_1 = 3 | X_0 = 2) \\ &= \Pr(H^4 < \infty | X_1 = 1, X_0 = 2) \Pr(X_1 = 1 | X_0 = 2) + \Pr(H^4 < \infty | X_1 = 3, X_0 = 2) \Pr(X_1 = 3 | X_0 = 2) \end{aligned}$$

By the Markov property, this simplifies to

$$\begin{aligned} h_2^4 &= \Pr(H^4 < \infty | X_1 = 1) \Pr(X_1 = 1 | X_0 = 2) + \Pr(H^4 < \infty | X_1 = 3) \Pr(X_1 = 3 | X_0 = 2) \\ &= \alpha h_1^4 + (1-\alpha) h_3^4 = (1-\alpha) h_3^4, \end{aligned}$$

where the last equality holds because 1 *is an absorbing state.*
With a similar argument, we now determine

$$h_3^4 = \beta h_2^4 + (1-\beta) h_4^4 = \beta h_2^4 + (1-\beta).$$

Combining these, we get

$$h_2^4 = (1-\alpha)\left(\beta h_2^4 + (1-\beta)\right),$$

so that

$$h_2^4 = \frac{(1-\alpha)(1-\beta)}{1-(1-\alpha)\beta}.$$

Remark 6.42. *This answer has a suggestive form - it looks like the infinite sum of a geometric series.*

$$h_2^4 = (1-\alpha)(1-\beta) \sum_{k=0}^{\infty} \left((1-\alpha)\beta\right)^k$$

This makes sense - h_2^4 is the probability of the event that the chain starts in state 2 *and ends up in state* 4*, which we write as $\{2 \rightsquigarrow 4\}$. This event can be expressed as a disjoint union*

$$\begin{aligned} \{2 \rightsquigarrow 4\} &= \{(2 \to 3 \to 4)\} && \text{from 2 to 4 via 3} \\ &\quad \cup \{(2 \to 3 \to 2 \to 3 \to 4)\} && \text{once around the } 2 \to 3 \to 2 \text{ loop} \\ &\quad \cup \{(2 \to 3 \to 2 \to 3 \to 2 \to 3 \to 4)\} && \text{twice around the } 2 \to 3 \to 2 \text{ loop} \\ &\quad \cup \dots \\ &= \bigcup_{n=0}^{\infty} \{2 \to 3 \to [2 \leftrightarrow 3]^n \to 4\} \end{aligned}$$

Proposition 6.43. *If $A \subseteq \mathcal{E}$, the vector $h^A = (h_i^A)_{i \in \mathcal{E}}$ solves the system of linear equations*

$$h_i^A = \begin{cases} 1 & i \in A \\ \sum_{j \in \mathcal{E}} p_{ij} h_j^A & i \notin A \end{cases}.$$

Moreover, h^A is the minimal solution, in the sense that if $x = (x_i)_{i \in \mathcal{E}}$ is another non-negative solution, then $h_i^A \leq x_i$ for all i.

Proof. *Note that if $X_0 = i \in A$, then certainly $H^A = 0$, so that $h_i^A = 1$.*
Now if $i \notin A$, then necessarily $H^A \geq 1$, so that we can condition on X_1:

$$\Pr(H^A < \infty | X_0 = i) = \sum_{j \in E} \Pr(H^A < \infty | X_0 = i, X_1 = j) \Pr(X_1 = j | X_0 = i).$$

By the Markov property, this is just

$$\sum_{j\in E} \Pr(H^A < \infty | X_1 = j) \Pr(X_1 = j | X_0 = i) = \sum_{j\in E} \Pr(H^A < \infty | X_1 = j) p_{ij} = \sum_{j\in E} p_{ij} h_j^A.$$

We now show minimality. Suppose $x = (x_i)_{i\in\mathcal{E}}$ is another non-negative solution of the linear system. We begin by showing that

$$x_i \geq \Pr\left(\bigcup_{n=0}^{N} \{X_n \in A\} | X_0 = i \right), \qquad N \in \mathbf{N}_0.$$

Certainly this is true for $N = 0$, because if $i \in A$, both sides are 1 and if $i \notin A$, the right hand expression is 0 and $x_i \geq 0$.

Suppose now that the result holds for $N \geq 0$. The relation clearly holds for $i \in A$. For $i \notin A$, we use the law of total probability to write

$$\begin{aligned}
\Pr\left(\bigcup_{n=0}^{N+1} \{X_n \in A\} | X_0 = i \right) &= \sum_{j\in\mathcal{E}} \Pr\left(\bigcup_{n=0}^{N+1} \{X_n \in A\}, X_1 = j | X_0 = i \right) \\
&= \sum_{j\in\mathcal{E}} \Pr\left(\bigcup_{n=0}^{N+1} \{X_n \in A\} | X_1 = j, X_0 = i \right) p_{ij} \\
&= \sum_{j\in\mathcal{E}} \Pr\left(\bigcup_{n=1}^{N+1} \{X_n \in A\} | X_1 = j \right) p_{ij} \\
&= \sum_{j\in\mathcal{E}} \Pr\left(\bigcup_{n=0}^{N} \{X_n \in A\} | X_0 = j \right) p_{ij}. \\
&\leq \sum_{j\in\mathcal{E}} p_{ij} x_j = x_i.
\end{aligned}$$

Hence the result holds for all $N \in \mathbf{N}_0$. It must therefore hold for the limit:

$$x_i \geq \lim_{N\to\infty} \Pr\left(\bigcup_{n=0}^{N} \{X_n \in A\} | X_0 = i \right).$$

But since the unions considered form an increasing sequence, this gives that

$$x_i \geq \Pr\left(H^A < \infty | X_0 = i \right) = h_i^A.$$

Stationary Distributions

Thinking back to our initial genetic example, we saw that the chain tended to forget its initial state: whatever the chain's initial state, after a long enough time, the marginal probability of being in any of the four states converged to $\frac{1}{4}$. The random fluctuations induced by mutation at each time step mean that there is no hope that the chain will settle down to a a definite state, nevertheless, the *probability*

distribution of being in any state does converge to a limit. This is true of Markov chains more generally, subject only to mild conditions on the transition matrix. We will explore this behaviour informally in this final section.

Explicitly, recall from the genetic example that for a state $i \in \{A, G, C, T\}$, we found that

$$p_{ii}(n) = \frac{1}{4} + \frac{3}{4}\left(1 - \frac{4\alpha}{3}\right)^n,$$

so that for $i \neq j$, by symmetry of the mutation process,

$$p_{ij}(n) = \frac{1}{3}(1 - p_{ii}) = \frac{1}{4} - \frac{1}{4}\left(1 - \frac{4\alpha}{3}\right)^n.$$

In matrix form

$$P^n = \frac{1}{4}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} + \frac{1}{4}\left(1 - \frac{4\alpha}{3}\right)^n \begin{pmatrix} 3 & -1 & -1 & -1 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ -1 & -1 & -1 & 3 \end{pmatrix}.$$

In the earlier example, we started the chain in the state $X_0 = T$. More generally, we might not know the initial state, and so instead we have an initial distribution λ_0, say.

The distribution in the next generation is given by $\lambda_1 = \lambda_0 P$, and more generally the distribution in generation n is given by $\lambda_n = \lambda_0 P^n$. Since λ is a probability distribution, multiplication by the first matrix term just gives

$$\lambda_\infty = \begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}$$

so the only dependency on λ_0 is through the second matrix term, which vanishes as $n \to \infty$.

Note an interesting property of λ_∞ as defined above: $\lambda_\infty P = \lambda_\infty$. We say that λ_∞ is **stationary** for the chain (X_n).

Definition 6.44. *A vector $\pi = (\pi_j)_{j \in \mathcal{E}}$ is said to be a stationary distribution for the Markov chain (X_n) if*

1. *$\pi_j \geq 0$ for all $j \in \mathcal{E}$ and $\sum_{j \in \mathcal{E}} \pi_j = 1$ (so π is a probability distribution on $\mathcal{E}$)*
2. *$\pi P = \pi$.*

Proposition 6.45. *If the distribution of X_n is π and π is stationary for the chain (X_n), then the distribution of X_{n+1} is also π.*

Proof

$$\Pr(X_n = j) = \sum_{i \in \mathcal{E}} \Pr(X_n = j | X_{n-1} = i) \Pr(X_{n-1} = i) = \sum_{i \in \mathcal{E}} p_{ij} \pi_i.$$

But this is just $(\pi P)_j = \pi_j$*, since* π *is stationary for* (X_n)*.*

Proposition 6.46. *(Non-examinable - see Grimmett and Stirzaker) An irreducible chain has a stationary distribution if and only if all states are positive recurrent. In this case* $\pi_j = \frac{1}{\mu_j}$*, where* μ_j *is the mean recurrence time, hence the stationary distribution is unique.*

Informal justification.

We take as a hopefully plausible starting point the **Ergodic theorem** *(below). This says that almost surely, the long-run proportion of the time spent in state i is* $\frac{1}{\mu_i}$*.*

At every visit to state i, the chain has a probability p_{ij} *of moving to state j. This gives the long-run proportion of jumps from i to j as* $\frac{1}{\mu_i} p_{ij}$*. Then (at least when* $\mathcal{E}$ *is finite), the total long-run proportion of jumps into state j is then just* $\sum_{i \in \mathcal{E}} \frac{1}{\mu_i} p_{ij}$*.*

But this is just the long-run proportion of time spent in state j, which is $\frac{1}{\mu_j}$*, by the ergodic theorem. This says that the vector with entries* $\pi_j = \frac{1}{\mu_j}$ *is a stationary distribution for the chain.*

Proposition 6.47. *(Non-examinable - see Grimmett and Stirzaker) If* $(X_n)_{n \in \mathbf{N}_0}$ *is an irreducible, aperiodic Markov chain with stationary distribution* π*, then, for any initial distribution* λ *and for any* $j \in \mathcal{E}$*,*

$$\lim_{n \to \infty} \Pr(X_n = j) = \pi_j,$$

in particular, for all $i \in \mathcal{E}$*, the limiting probability*

$$\lim_{n \to \infty} \Pr(X_n = j | X_0 = i) = \pi_j$$

is independent of i.

Proposition 6.48. *The Ergodic theorem. (Non-examinable - see Grimmett and Stirzaker) Let* $(X_n)_{n \in \mathbf{N}_0}$ *be an irreducible Markov chain. For any state* $i \in \mathcal{E}$*, let*

$$V_i(n) = \sum_{r=0}^{n} I(X_r = i)$$

count the number of visits to i before time n. Then for any initial distribution and and $i \in \mathcal{E}$*,*

$$\Pr\left(\frac{V_i(n)}{n} \to \pi_i \text{ as } n \to \infty \right) = 1,$$

i.e. $\frac{V_i(n)}{n}$ *converges almost surely to* π_i*.*

Example 6.49. *The two state chain. We ignore degenerate cases for which $\alpha = \beta = 0$ and $\alpha = \beta = 0$. We solve for $\pi P = \pi$*

$$\begin{pmatrix} \pi_1 & \pi_2 \end{pmatrix} \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix} = \begin{pmatrix} \pi_1 & \pi_2 \end{pmatrix},$$

so that

$$\pi_1(1-\alpha) + \pi_2\beta = \pi_1,$$

giving

$$\pi_2\beta = \pi_1\alpha.$$

For a normalizable distribution, we must have $\pi_2 = 1 - \pi_1$, so that

$$\pi_1 = \frac{\beta}{\alpha+\beta} \qquad \pi_2 = \frac{\alpha}{\alpha+\beta},$$

limiting probabilities that we have already established by explicitly computing the n-step transition probabilities.

Example 6.50. *Let $\mathcal{E} = \{1, 2, \ldots N\}$, where N is odd. Suppose $(X_n)_{n\in\mathbf{N}_0}$ is the symmetrical random walk on the polygon with vertices labelled by elements of $\mathcal{E}$. The condition for stationarity in this case is*

$$\pi_i = \frac{1}{2}\pi_{i-1} + \frac{1}{2}\pi_{i+1}.$$

It is immediate from symmetry that $\pi_i = \frac{1}{N}$, and $p_{ii}(n) \to \frac{1}{N}$ as $n \to \infty$.

(What happens when N is even?)

Example 6.51. *Consider Figure 6.8. Note that this is just a random walk on the graph defined by the five vertices: the chain chooses uniformly from the options available to it at each point. To find the stationary distribution, we have five unknowns to solve for. However, the problem has some symmetry: clearly $\pi_1 = \pi_2$ and $\pi_3 = \pi_4$, so we have only three unknowns, and, since the entries of π sum to one, the problem is simpler still.*

We solve for $\pi P = \pi$. From the first column:

$$\frac{1}{2}\pi_2 + \frac{1}{3}\pi_4 = \pi_1,$$

so that, since $\pi_2 = \pi_1$, we have $\frac{1}{2}\pi_2 = \frac{1}{3}\pi_4$.
Now from the fifth column:

$$\frac{1}{3}\pi_3 + \frac{1}{3}\pi_4 = \pi_5,$$

so that, since $\pi_3 = \pi_4$, we have $\frac{2}{3}\pi_4 = \pi_5$.

We now have all we need. The vector $\pi = \left(\frac{2}{3}\pi_4, \frac{2}{3}\pi_4, \pi_4, \pi_4, \frac{2}{3}\pi_4\right)$. *Hence*

$$\pi_4\left(\frac{2}{3}+\frac{2}{3}+1+1+\frac{2}{3}\right) = 1,$$

so that $\pi_4 = \frac{1}{4}$ *and*

$$\pi = \left(\frac{1}{6}, \frac{1}{6}, \frac{1}{4}, \frac{1}{4}, \frac{1}{6}\right).$$

This is in fact an instance of a more general result. For a symmetrical random walk on a finite graph, if the state i *is connected to* d_i *other states, then* $\pi_i = \frac{d_i}{\sum_{j\in\mathcal{E}} d_j}$.

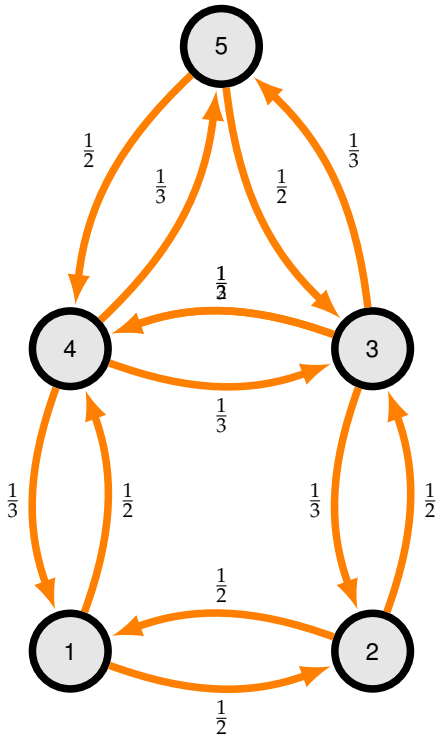


Figure 6.8: A random walk on a graph with five vertices.