

STATS 125: Probability and its Applications.

Marie Fitch

Azam Asanjarani

Semester 1, 2019

This lecture workbook contains gapped notes and optional end of chapter exercises. We will assume that you have the relevant pages with you at lectures. A filled-in version of these notes will be available on Canvas at the conclusion of lectures on each chapter. Brief solutions to the optional exercises will also be available on Canvas.

Thanks to previous STATS125 lecturers for their contributions to the previous versions of the workbook on which this workbook is based.

*Cover photo: Māngere Mountain (Te Pane o Mataaho), NZ.
Thanks to Nigel Fitch*

Contents

1	Introduction	7
2	Sample spaces and events	11
2.1	Sample spaces	11
2.2	Events	12
2.3	Partitions	20
2.4	Vocabulary and symbols	22
2.5	Exercises	23
3	Probability	25
3.1	Probability measure	25
3.2	Probability rules	26
3.3	Equally likely outcomes	32
3.4	Exercises	35
4	Conditional probability, Bayes' theorem and independence	37
4.1	Conditional probability	37
4.2	Bayes' Theorem	45
4.3	Independence	49
4.4	Equally likely outcomes and independence	53
4.5	Exercises	54
5	Discrete random variables, expectation and variance	61
5.1	Random variables	61
5.2	Probability mass function	62
5.3	The cumulative distribution function	64

5.4	Expected value	68
5.5	Expected value of a function of a random variable	70
5.6	Expected value of sums and differences of random variables	72
5.7	Variance	73
5.8	Exercises	79
6	Discrete distributions	85
6.1	Bernoulli random variables	86
6.2	Binomial distribution	88
6.3	Geometric distribution	95
6.4	Negative Binomial distribution	98
6.5	Hypergeometric distribution	102
6.6	Discrete Uniform distribution	107
6.7	Poisson distribution	108
6.8	Exercises	111
7	Joint and conditional distributions	117
7.1	Joint distributions	117
7.2	Independent random variables	124
7.3	Conditional distributions	127
7.4	Testing for independence	131
7.5	Relations between distributions	135
7.6	Exercises	140
8	Covariance and conditional expectation	141
8.1	Covariance	141
8.2	Correlation	146

8.3	Conditional expectation	148
8.4	Prediction	152
8.5	Conditional expectation as a random variable	157
8.6	Probability, statistics and data	162
8.7	Exercises	167
9	Introduction to Markov chains	169
9.1	Examples	169
9.2	The Markov property	172
9.3	Transition matrices	173
9.4	Sample path behaviour and n -step transition probabilities	176
9.5	Exercises	182
10	Equilibrium and limiting distributions	183
10.1	Equilibrium distributions and the Full Balance Equations	183
10.2	Detailed Balance Equations	191
10.3	Limiting distributions	197
10.4	Qualitative features of Markov chains	201
10.5	When are the equilibrium and limiting distributions equal?	204
10.6	Optional section: Equilibrium distributions, limiting distributions, and transition matrices	206
10.7	Exercises	208
11	Hitting/reaching probabilities and times	213
11.1	Hitting probabilities	213
11.2	Expected hitting times	216
11.3	Exercises	220

1 Introduction

Probability forms the basis of statistics and is used in many other fields, including biological sciences, mathematical physics, and commerce. Probability is the study of randomness or uncertainty. Whether or not you believe that there is anything truly random in the universe (this is a philosophical question), there are certainly many things that are unpredictable. If this were not the case, we would not have things like lotteries, casinos and insurance, and no one would pay more to watch sport *live* on TV!

In order to determine what price you should pay for your insurance, insurance companies estimate the probability or likelihood of the “bad event” (that you are to be insured against) happening within a certain period of time. Roughly speaking, the more information they have about you, the better their estimate will be. Of course the bad event could happen to you at any time, but the insurance company relies on the fact that, assuming its estimates are reasonable, it is very unlikely that they will have to pay out money to a large proportion of their customers in the near future. That is to say that they benefit from having a large number of customers whose “risk” they have assessed.

Casinos make their money in a similar way, except that they don’t really have to “estimate” the risk in most of the games. For example, if you roll a well-constructed/fair 6-sided die, you “know” what the possible outcomes are and the likelihood of each possible outcome occurring. A well designed lottery is the ultimate in money-making arrangements. For example, the NZ lottery can essentially not lose money. The basic idea of the lottery is to collect money from lottery players, keep some of it, and redistribute the remainder among the players!

Randomness relates to the idea that we are unable to exactly predict the outcome of a particular event.

e.g., coin tossing; each time the coin is tossed, we know that it will be heads or tails but we cannot be sure which one of these will occur

Probability relates to how likely a particular outcome is.

e.g., most people agree that the chance of getting heads when you toss a fair coin is 0.5

A probability is a number between 0 and 1 representing how likely it is that an event will occur. What does this actually mean?

In terms of coin tossing, this is relatively easy to comprehend: if we toss the coin a whole lot of times, we expect about half of the tosses to be heads.

This is the **frequentist** view of probability (“frequentist” simply means “based on frequencies”), e.g. the probability of getting tails when a coin is tossed once is based on how frequently (i.e. how often) you get tails when you toss the coin a large number of times.

$$\text{e.g., } \frac{\text{number of tails}}{\text{number of coin tosses}}$$

This assumes that all coin tosses are performed under basically the same conditions. This is true for coin tossing, but what about other things that we assign probabilities to?

e.g., The probability our football team wins tomorrow is 0.6

Is this probability based on the proportion of times our team wins over a long time?

Suppose we know that it has beaten the other team only 10 times in the past 100 meetings. Then perhaps 0.1 would be a good estimate of the probability of a win tomorrow.

However, suppose our team has won 16 of its last 20 games. We might think a probability of **0.8** of a win tomorrow might be more appropriate

In reality there is no “right way” to answer this question, and we will probably choose a number between zero and one based on many things including injuries to key players, the weather, etc. The number we choose simply reflects how **confident** we are that the team will win tomorrow. Even though 10 fans may have access to the same information, they may come up with vastly different numbers for the probability of winning tomorrow.

This is a **subjectivist** view of probability.

Although different interpretations of probability lead to quite different ways of thinking about things, the language of probability is the same in each case.

Probabilities can be:

1. Frequentist (based on frequencies),

$$\text{e.g.} \quad \frac{\text{number of times event occurs}}{\text{number of opportunities for event to occur}}$$

or

2. Subjective: probability represents a person's **degree of belief** that an event will occur, e.g., I think there is an 80% chance we will win tomorrow

When attempting to answer a question of the form “What is the probability of ??” we first need to ensure that the question is well defined and that it is actually possible to get an answer. For example, “What is the probability of orange?” is not a well defined question. Is it asking about the fruit or the colour? What is the context? To address this kind of problem, we begin this course by setting up the framework and rules on which the study of probability is based.

Regardless of how we obtain probabilities, we always combine and manipulate them according to the same rules.

Some texts:- While these course notes are designed to be sufficient in themselves if you wish to consult other texts the following are suggested. Note that all include some material that is not a part of this course and most do not include all the material covered in this course. Those that are available as e-books are also listed separately in Canvas.

Bertsekas, D.P. and Tsitsiklis, J.N. (2002) *Introduction to probability*. Athena Scientific. QA273 .B47 2002

Chung, K.L. (2003) *Elementary probability theory : with stochastic processes and an introduction to mathematical finance*. 4th ed. Springer. 519.23 C55 2003

The e-resource and earlier editions are also useful.

Grimmett, G. and Welsh, D. (1986) *Probability: An Introduction*. Oxford University Press. 519.2 G86p

The e-resource 2nd edition is also useful.

Grinstead, C.M. and Snell, J.L. (1997) *Introduction to probability*. American Mathematical Society. 519.2 G868

Haigh, J. (2002) *Probability models*. Springer. 519.2 H149p

The e-resource 2nd edition is also useful.

Pitman, J. (1993) *Probability*. Springer. 519.2 P685

Ross, S.M. (2007) *Introduction to probability models*. 8th ed. Harcourt/Academic Press 519.2 R82 2007

The earlier editions and e-resource 10th edition are also useful.

Some more advanced texts that you might like to look at:-

Feller, W. (1957) *An Introduction to Probability theory and its applications*. Vol. 1. Wiley. 519.2 F31 1957

Whittle, P. (2000) *Probability via Expectation*. Springer. 519.2 W627 2000

Williams, D. (2001) *Weighing the Odds: A Course in Probability and Statistics*. Cambridge University Press. 519.2 W72w

Some historical accounts:-

Gani, J. (ed.) (1986) *The Craft of Probabilistic Modelling: A Collection of Personal Accounts*. Springer. 519.2 C885

This includes autobiographical essays by two very well-known New Zealanders – David Vere-Jones and Peter Whittle.

Heyde, C.C. and Seneta, E. (eds.) (2001) *Statisticians of the centuries*. Springer. 519.50922 H61

2 Sample spaces and events

By the end of this chapter you should be able to:

- correctly use the vocabulary and symbols listed on page 22
- draw and interpret Venn diagrams for 2 or 3 events
- identify mutually exclusive events

2.1 Sample spaces

Definition: A **random experiment** is an experiment whose outcome is not known until it is observed.

Definition: A **sample space**, Ω , is the set of possible outcomes of a random experiment.

Every possible outcome should be listed *once and once only*.

Definition: A **sample point**, ω , is an element of the sample space.

For example, if the sample space is $\Omega = \{\omega_1, \omega_2, \omega_3\}$, then each ω_i is a sample point.

To begin with, we are only going to consider **finite** sample spaces (these are spaces which contain a finite number of elements).

Example 2.1. Toss a coin and observe the result. There are only two possible outcomes: heads or tails. We can define the sample space as $\Omega = \{H, T\}$

Ω contains all possible outcomes of our coin tossing experiment. Each of those outcomes is a sample point.

e.g., the outcome “heads”, or “ H ”, is a sample point within the sample space, Ω . The outcome “tails”, or “ T ”, is another sample point within Ω .

Example 2.2. Roll a six-sided die and observe the result. There are six possible outcomes: 1, 2, 3, 4, 5, 6.

$\Omega = \{1, 2, 3, 4, 5, 6\}$. Each of the outcomes is a sample point

2.2 Events

2.2.1 Events and Venn diagrams

Definition: An **event** is a subset of the sample space. That is, any collection of outcomes forms an event.

Example 2.2 *continued.* Roll a six-sided die. Sample space: $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Let A be the event that the observed result is an even number.

We write $A = \text{"roll an even number"}$

Then $A = \{2, 4, 6\}$

A is a subset of Ω , as in the definition. We write $A \subset \Omega$, sometimes written $A \subseteq \Omega$.

Definition: Event A **occurs** if we observe an outcome that is a member of the set A .

Note: Ω is a subset of itself, so Ω is an event. Because Ω includes all possible outcomes of the experiment, event Ω occurs every time the experiment is performed.

The set containing no outcomes, $\emptyset = \{\}$, is also a subset of Ω . This is called the empty set.

Example 2.3. Draw three cards from a deck and observe only the colour of each card. Each card can be black or red.

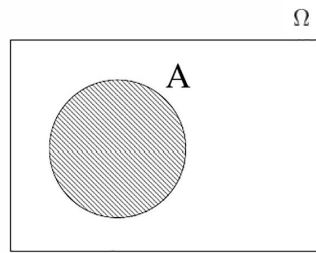
We are interested in the event where we observe less than two red cards.

Sample space: $\Omega = \{RRR, RRB, RBR, BRR, RBB, BRB, BBR, BBB\}$

(Note that this assumes that the cards are drawn one at a time and that the colours are listed in the order that the three cards are drawn.)

Event $A = \text{"less than two red cards"} = \{RBB, BRB, BBR, BBB\}$

Venn diagrams provide a graphical representation of our set notation, and can be useful for visualizing the relationship between events. An event $A \subset \Omega$ can be represented graphically in a Venn diagram as:

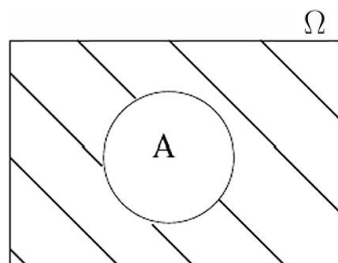


2.2.2 Complement of an event

Definition: The **complement** of event A is written A^c (or sometimes \overline{A}), and is given by

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

That is, A^c is the event “not A ”, i.e. A does not occur. In a Venn diagram (A^c shaded) this is shown as:



Example 2.4. Toss a coin, $\Omega = \{H, T\}$, and let $A = \text{“heads”} = \{H\}$. Then $A^c = \{T\}$.

2.2.3 Union of events

Example 2.5. Pick a person in class.

Sample space $= \Omega = \{\text{all people in class}\}$.

Let event $A = \text{“person is taller than 1.7m”}$.

Let event $B = \text{“person is female”}$.

Then event “person is either taller than 1.7m or a female or both” is the **union** of events A and B and is denoted by $A \cup B$.

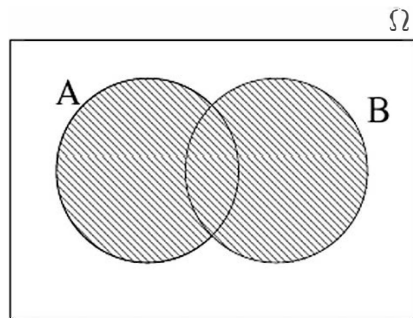
Definition: Let A and B be events on the same sample space Ω , so $A \subset \Omega, B \subset \Omega$. The **union** of events A and B is written $A \cup B$ and is defined to be the event

$$A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B \text{ or both}\}.$$

(Read this as “the set of points ω in Ω such that ω is a point in A , or ω is a point in B , or both”.) Often (when it is clear what sample space our points belong to) we drop the sample space Ω from the notation and just write

$$A \cup B = \{\omega : \omega \in A \text{ or } \omega \in B \text{ or both}\}.$$

Think of $A \cup B$ as A or B or both. On a Venn diagram, we show $A \cup B$ as follows:



Example 2.6. Toss two coins and observe the outcomes.

We can represent the sample space as: $\Omega = \{HH, HT, TH, TT\}$.

Then the event “two heads observed” is $A = \{HH\}$

and the event “two tails observed” is $B = \{TT\}$

Then $A \cup B = \{HH\} \cup \{TT\} = \{HH, TT\}$

= “two heads or two tails observed”

2.2.4 Intersection of events

Example 2.7. Pick a person in class.

Sample space $\Omega = \{\text{all people in class}\}$.

Let event $A = \text{“person has black hair”}$.

Let event $B = \text{“person is male”}$.

Then event “person has black hair *and* is male” is the **intersection** of events A and B is written $A \cap B$.

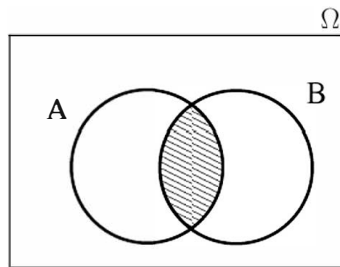
Definition: Let A and B be events on the same sample space Ω , so $A \subset \Omega, B \subset \Omega$. The **intersection** of events A and B is denoted by $A \cap B$ and is given by

$$A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}$$

or, equivalently

$$A \cap B = \{\omega : \omega \in A \text{ and } \omega \in B\}.$$

Think of $A \cap B$ as “*both A and B*”. On a Venn diagram, we show $A \cap B$ as follows:



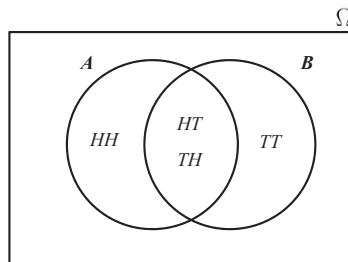
Example 2.8. Toss two coins and observe the outcomes.

We can represent the sample space as: $\Omega = \{HH, HT, TH, TT\}$.

The event “at least one head is observed” is $A = \{HH, HT, TH\}$

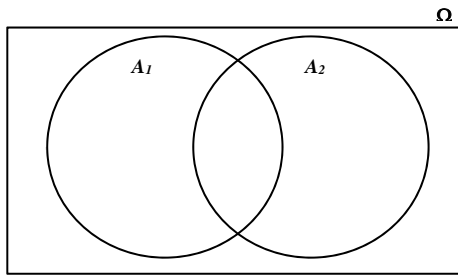
and the event “at least one tail is observed” is $B = \{HT, TH, TT\}$

$A \cap B = \{HT, TH\}$.

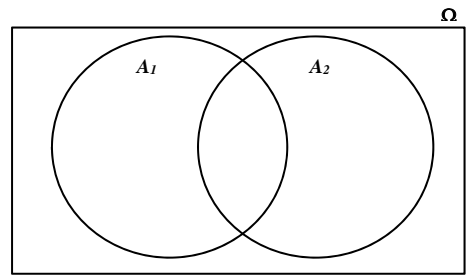


2.2.5 Intersections, unions and complements

Example 2.9. On the Venn diagrams below shade (a) $A_1^c \cap A_2^c$ and (b) $A_1^c \cup A_2^c$, then in each case state how else you could describe what you have shaded.

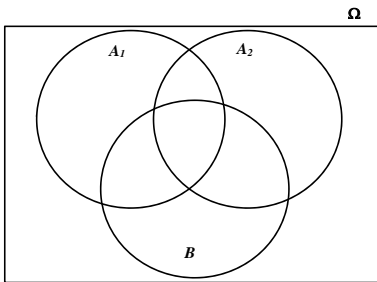


(a) $A_1^c \cap A_2^c = (A_1 \cup A_2)^c$

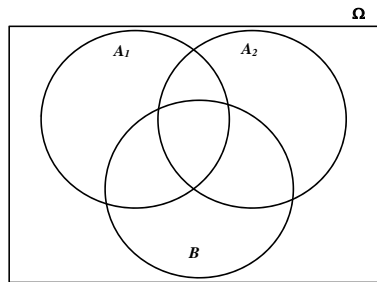


(b) $A_1^c \cup A_2^c = (A_1 \cap A_2)^c$

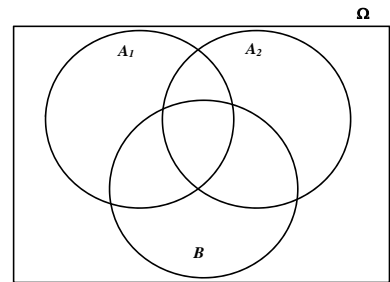
Example 2.10. Shade each Venn diagram to match the description below it.



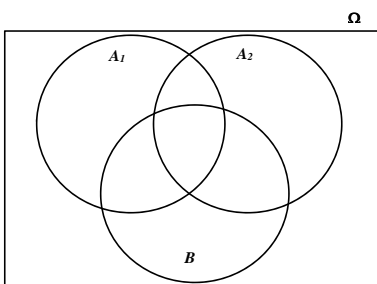
(a) $B \cap A_1$



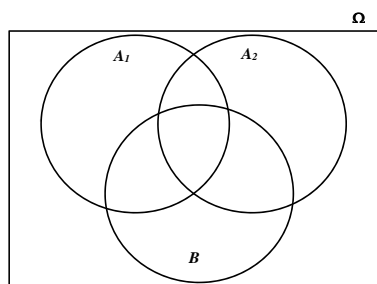
(b) $B \cap A_2$



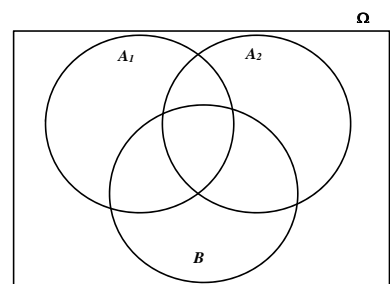
(c) $(B \cap A_1) \cup (B \cap A_2)$



(d) $B \cup A_1$



(e) $B \cup A_2$

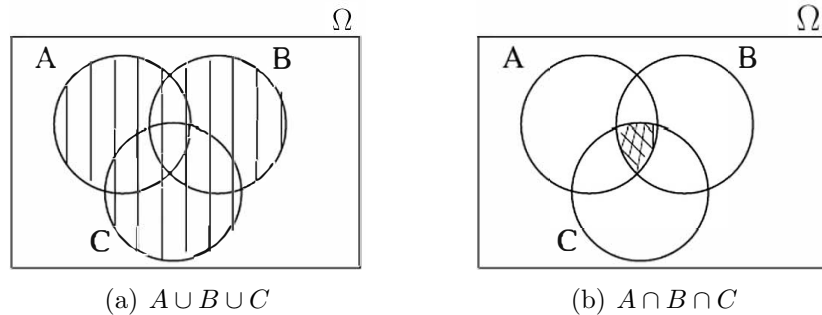


(f) $(B \cup A_1) \cap (B \cup A_2)$

2.2.6 Three events:

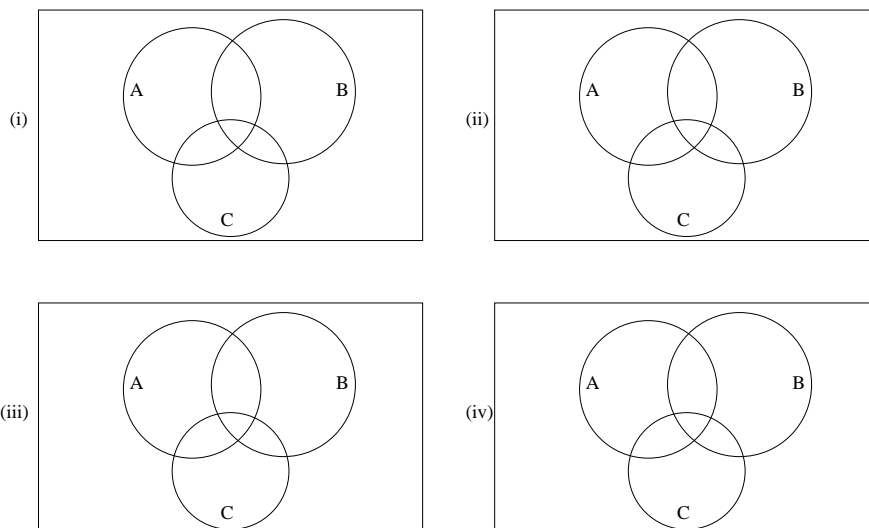
Venn diagrams are generally useful for up to 3 events.

For example:



When we want to combine events using both unions and intersections we need to be careful.

Example 2.11. (a) On the Venn diagrams below shade (i) $A \cup B$, (ii) $(A \cup B) \cap C$, (iii) $B \cap C$ and (iv) $A \cup (B \cap C)$.



(b) What do you notice about $(A \cup B) \cap C$ and $A \cup (B \cap C)$?

Thus $(A \cup B) \cap C \neq A \cup (B \cap C)$, so that $A \cup B \cap C$ has no meaning without the brackets.

Example 2.12. (a) Look back to Example 2.10, how else could you describe the shading in (c)?

Complete the statement: $(B \cup A_1) \cap (B \cup A_2) = B \cup (A_1 \cap A_2)$

(b) Look back to Example 2.10, how else could you describe the shading in (f)?

Complete the statement: $(B \cap A_1) \cup (B \cap A_2) = B \cap (A_1 \cup A_2)$

2.2.7 More than three events

It is useful to introduce some notation to deal with large numbers of events. We'll use the notation

$$\bigcup_{i=1}^k A_i := A_1 \cup A_2 \cup \cdots \cup A_k, \quad \text{and} \quad \bigcap_{i=1}^k A_i := A_1 \cap A_2 \cap \cdots \cap A_k,$$

for example, $\bigcup_{i=1}^3 A_i = A_1 \cup A_2 \cup A_3$.

Using this notation we can generalise the relationships we observed in the previous section.

For any collection of events A_i , the following set relations hold:

$$\begin{aligned} \left(\bigcap_i A_i \right)^c &= \bigcup_i A_i^c & \left(\bigcup_i A_i \right)^c &= \bigcap_i A_i^c \\ B \cap \left(\bigcup_i A_i \right) &= \bigcup_i (B \cap A_i) & B \cup \left(\bigcap_i A_i \right) &= \bigcap_i (B \cup A_i) \end{aligned}$$

The Venn diagrams from the previous section are great for convincing ourselves that in the case of two or three events these relationships hold but they do not constitute a proof.

There are a few different ways to show that two sets A and B are equal. The general idea is to show that $A \subset B$ and $B \subset A$. That is, if $\omega \in A$, then $\omega \in B$ and vice versa. It's often just as easy to do both at the same time. For example

$$\begin{aligned} \left(\bigcap_i A_i \right)^c &= \{ \omega \in \Omega : \omega \notin \bigcap_i A_i \} = \{ \omega \in \Omega : \omega \notin A_i \text{ for some } i \} \\ &= \{ \omega \in \Omega : \omega \in A_i^c \text{ for some } i \} = \{ \omega \in \Omega : \omega \in \bigcup_i A_i^c \} = \bigcup_i A_i^c, \end{aligned}$$

or equivalently, using the mathematics notation “ \iff ” (=“if and only if”), “ \exists ” (=“there exists”) and “ \cdot ” (=“such that”),

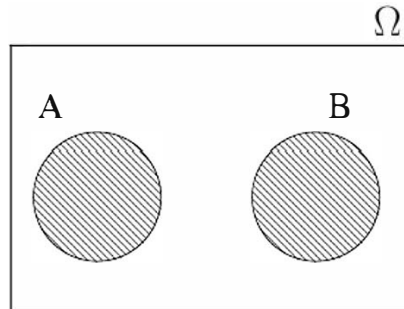
$$\omega \in \left(\bigcap_i A_i \right)^c \iff \omega \notin \bigcap_i A_i \iff \exists i : \omega \notin A_i \iff \exists i : \omega \in A_i^c \iff \omega \in \bigcup_i A_i^c.$$

Challenge Exercise (*non-examinable*): Prove the other three set relations.

2.3 Partitions

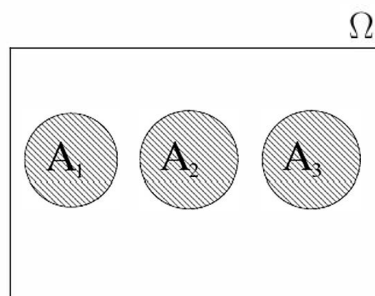
Definition: Two events A and B are **mutually exclusive**, or **disjoint**, if $A \cap B = \emptyset$.

This means events A and B cannot happen together. If A happens, it excludes B from happening, and vice-versa.



Note: A and A^c are mutually exclusive.

Definition: Any number of events A_1, A_2, \dots, A_k are **mutually exclusive** if every pair of the events is mutually exclusive: i.e., $A_i \cap A_j = \emptyset$ for all i, j with $i \neq j$.

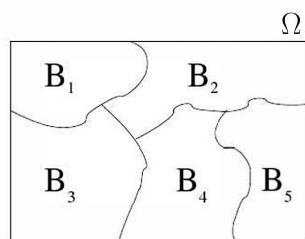


Definition: A **partition** of B is a collection of mutually exclusive events whose union is B .

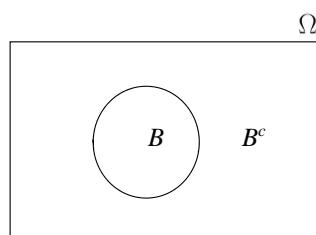
That is, sets B_1, B_2, \dots, B_k form a partition of B if

$$B_i \cap B_j = \emptyset \text{ for all } i, j \text{ with } i \neq j, \\ \text{and } \bigcup_{i=1}^k B_i = B_1 \cup B_2 \cup \dots \cup B_k = B.$$

Examples of partitions of Ω

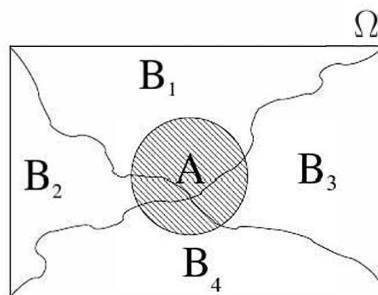


B_1, \dots, B_5 partition Ω



B and B^c partition Ω

Note that if B_1, \dots, B_k form a partition of B , **and** $A \subset B$ then $(A \cap B_1), \dots, (A \cap B_k)$ form a partition of A .



We will see that this is very useful for finding the probability of event A .

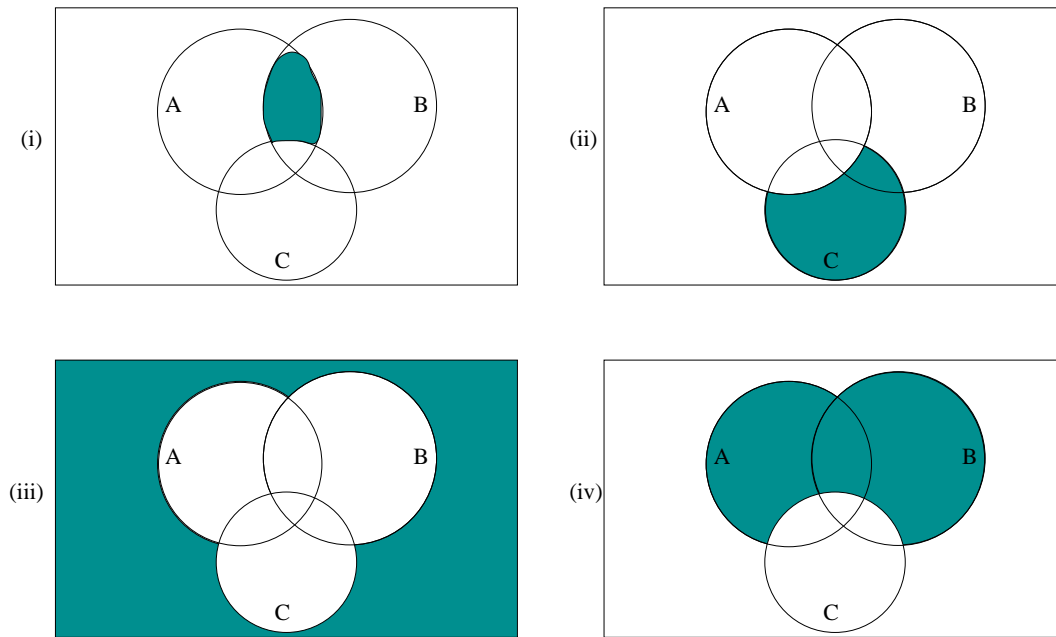
2.4 Vocabulary and symbols

The following terms and symbols have been defined and used in this chapter. You should now be able to use them appropriately.

Term	meaning	Symbol
random experiment		
sample space	the set of all possible outcomes	Ω
sample point	one possible outcome	$\omega_1, \omega_2, \dots$
event	a set of possible outcomes	A (or any other capital letter)
subset		\subset
empty set		\emptyset
union	or ('at least one of')	\cup also $\bigcup_{i=1}^k A_i$
intersection	and (both, 'all of')	\cap also $\bigcap_{i=1}^k A_i$
complement	not	A^c
mutually exclusive	$A \cap B = \emptyset$	
partition of A	mutually exclusive and together make-up A (‘jigsaw puzzle pieces’)	

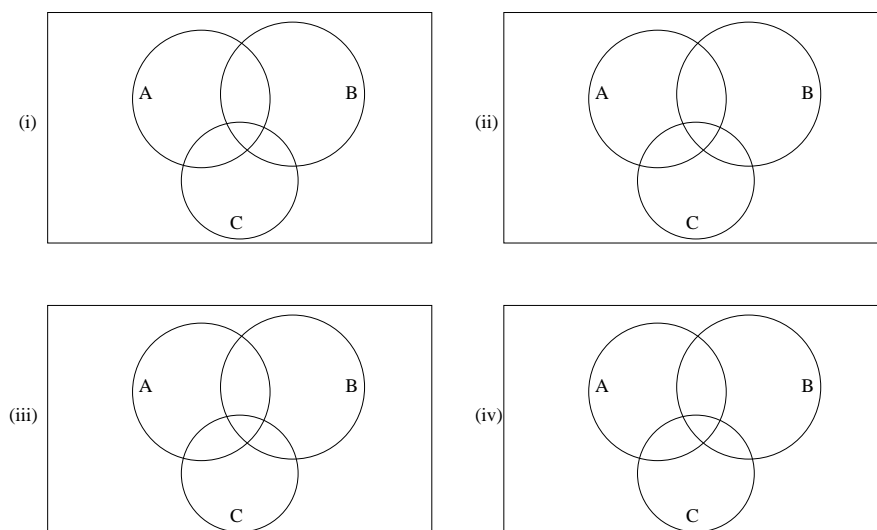
2.5 Exercises

2.5.1 Give the set notation for the shaded regions in the following Venn diagrams.



2.5.2 Use shading in the Venn diagrams below to identify the following sets:

- (i) $(A \cup B \cup C)^c$
- (ii) $B \cap C$
- (iii) $(A \cup B \cup C)^c \cup (B \cap C)$
- (iv) $(A \cup B \cup C) \cap (B \cap C)^c$



2.5.3 Deoxyribonucleic acid (DNA) consists of the bases A, T, C, and G. At each point in our genome, we each have one of these bases on the ‘sense’ strand of our DNA. Suppose that I randomly select two places in the genome at which to investigate base identity. That is, I want to find out which of the four bases are present at each of two locations. (Although not strictly true, lets assume that each letter is equally likely to occur).

- (a) What is the sample space?
- (b) Suppose event B = “At least one **T** is in the pair”. List the sample points in B .
- (c) How many points in the sample space exist for which the two bases have different letters?
- (d) Let B_1 = “first letter is **A** or **T**” and B_2 = “second letter is **C** or **G**”. Do B_1 and B_2 form a partition of Ω ? Why or why not?

2.5.4 Suppose we toss a coin three times.

- (a) What is the sample space, Ω , for this experiment?
- (b) List the sample points in each of the following events:
 - (i) A = “at least two heads”
 - (ii) B = “at least one tail”
 - (iii) $A \cap B$
 - (iv) A^c
 - (v) $A \cup B^c$
- (c) Draw a Venn diagram to represent the following events in the sample space:
 - (i) Event A from part (b) above.
 - (ii) Event B from part (b) above.
 - (iii) $A \cap B$

3 Probability

By the end of this chapter you should be able to:

- make connections between visual and symbolic representations of probabilities
- use probability rules to solve problems
- distinguish between situations when events are and are not equally likely.

3.1 Probability measure

Definition: A **probability measure** \mathbb{P} on a sample space Ω is a function that maps subsets of Ω to numbers in $[0, 1]$ (i.e. between 0 and 1, inclusive) such that:

- $\mathbb{P}(\emptyset) = 0$
- $\mathbb{P}(\Omega) = 1$
- If A_1, A_2, \dots are a collection of mutually exclusive (disjoint) events then

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

For a *discrete* sample space Ω (discrete Ω means any finite set or set that can be listed as $\Omega = \{s_1, s_2, s_3, \dots\}$), the third part of the definition above says that the probability of an event A is the sum of the probabilities of the individual outcomes in A .

This is because the events $A_i = \{\omega_i\}$, $i = 1, 2, 3, \dots$ are mutually exclusive (elements of Ω are listed only once).

Example 3.1. Toss a coin. Then $\Omega = \{\omega_1, \omega_2\}$, where the two possible outcomes are $\omega_1 = H$ and $\omega_2 = T$.

$$\mathbb{P}(\{\omega_1\}) = 0.5, \quad \mathbb{P}(\{\omega_2\}) = 0.5$$

$$\mathbb{P}(\Omega) = \mathbb{P}(\{\omega_1\}) + \mathbb{P}(\{\omega_2\}) = 0.5 + 0.5 = 1$$

Example 3.2. Two children in a family. Each child is either male or female. In a random family containing two children:

Experiment 1: observe whether each child is a boy or a girl

Then $\Omega = \{MM, MF, FM, FF\}$, and we model the experiment by declaring that each outcome is equally likely.

Let $A = \text{"one boy"}$.

Then $A = \{MF, FM\}$ and $\mathbb{P}(A) = 0.25 + 0.25 = 0.5$.

Experiment 2: observe the number of girls

Then $\Omega = \{0, 1, 2\}$

and

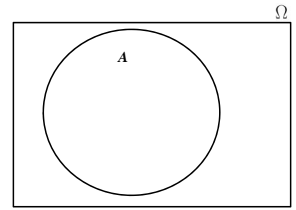
$\mathbb{P}(\{0\}) = 0.25, \mathbb{P}(\{1\}) = 0.5, \mathbb{P}(\{2\}) = 0.25$.

Let $B = \text{"at least one girl"}$

Then $B = \{1, 2\}$ and $\mathbb{P}(B) = 0.50 + 0.25 = 0.75$

3.2 Probability rules

1. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ for any event A .



2.

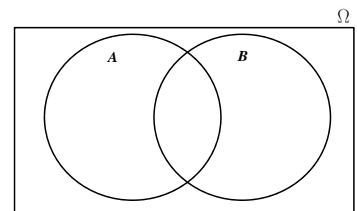
Theorem 3.1. The Partition Theorem

If B_1, B_2, \dots, B_m form a partition of Ω , then for any event A ,

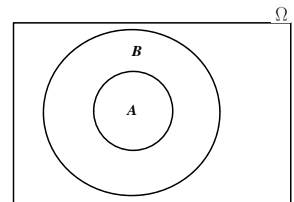
$$\mathbb{P}(A) = \sum_{i=1}^m \mathbb{P}(A \cap B_i).$$

Special case:

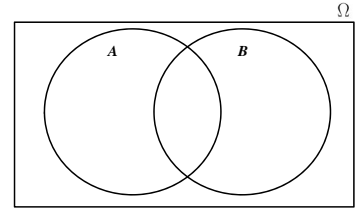
$\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c)$ for any events $A, B \subset \Omega$.



3. If $A \subset B$ then $\mathbb{P}(A) \leq \mathbb{P}(B)$



4. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$ for any events $A, B \subset \Omega$.



1. From Chapter 2.3 we know that A and A^c partition Ω , i.e. $A \cup A^c = \Omega$ and $A \cap A^c = \emptyset$.
Therefore:

$$\begin{aligned} \mathbb{P}(A \cup A^c) &= \mathbb{P}(\Omega) = 1 && \text{by definition} \\ \text{and } \mathbb{P}(A \cup A^c) &= \mathbb{P}(A) + \mathbb{P}(A^c) && \text{by definition} \\ \Leftrightarrow \mathbb{P}(A) + \mathbb{P}(A^c) &= 1 \\ \Leftrightarrow \mathbb{P}(A^c) &= 1 - \mathbb{P}(A) \end{aligned}$$

2. **Special case:**

From Chapter 2.3 we know that B and B^c partition Ω and therefore partition A .
Thus

$$\begin{aligned} \mathbb{P}(A) &= \mathbb{P}((A \cap B) \cup (A \cap B^c)) \\ &= \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c) && \text{by definition} \end{aligned}$$

Theorem 3.1 is proved in a similar manner.

3.

$$\begin{aligned} \mathbb{P}(B) &= \mathbb{P}(A \cap B) + \mathbb{P}(A^c \cap B) && \text{by the partition theorem} \\ &= \mathbb{P}(A) + \mathbb{P}(A^c \cap B) && \text{since } A \subset B \\ \Rightarrow \mathbb{P}(A) &\leq \mathbb{P}(B) && \text{since } \mathbb{P}(A^c \cap B) \geq 0 \text{ by definition} \end{aligned}$$

4.

$$\begin{aligned} \mathbb{P}(A \cup B) &= \mathbb{P}(A \cup (A \cap B^c)) && \text{from chapter 2} \\ &= \mathbb{P}(A) + \mathbb{P}(A \cap B^c) && \text{by definition} \\ &= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) && \text{by the partition theorem} \end{aligned}$$

3.2.1 Visual Representations

For simple problems a table (of counts of or probabilities) or a Venn diagram can be a useful aid to understand what is going on.

Exercise Complete the table the relevant probabilities using symbolic probability notation.

	A	A^c	Total
B_1	$\mathbb{P}(B_1 \cap A)$	$\mathbb{P}(B_1 \cap A^c)$	$\mathbb{P}(B_1)$
B_2	$\mathbb{P}(B_2 \cap A)$	$\mathbb{P}(B_2 \cap A^c)$	$\mathbb{P}(B_2)$
B_3	$\mathbb{P}(B_3 \cap A)$	$\mathbb{P}(B_3 \cap A^c)$	$\mathbb{P}(B_3)$
Total	$\mathbb{P}(A)$	$\mathbb{P}(A^c)$	$\mathbb{P}(\Omega) = 1$

Sketch a Venn diagram to show the same events. Note that B_1 , B_2 , and B_3 form a partition of S .

3.2.2 Events and Probabilities

Make sure you are very clear when you are talking about events and when you are talking about probabilities.

* Events are sets so use set notation like \cup and \cap .

* Probabilities are numbers so use $+$ and \times .



Example 3.3. For each of the following pairs of statements circle the correct statement and then write it out in words.

1. $A \cup B$ or $A + B$

$A \cup B$ at least one of A or B occurs

2. $\mathbb{P}(A \cup B)$ or $\mathbb{P}(A) \cup \mathbb{P}(B)$

$\mathbb{P}(A \cup B)$ The probability that at least one of A or B occurs

3. $\mathbb{P}(A + B)$ or $\mathbb{P}(A) + \mathbb{P}(B)$

$\mathbb{P}(A) + \mathbb{P}(B)$ the probability that A occurs plus the probability that B occurs

4. $A \times B$ or $A \cap B$

$A \cap B$ both A and B occur

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

$\mathbb{P}(A \cap B)$ the probability that both A and B occur

In many examples, such as the one below we use the phrase “a randomly chosen” item. When we use this phrase we are always talking about the specific situation where each item has the exact same chance of being selected.

Example 3.4. Gender, obesity and perceptions

Obesity is becoming a major problem in western culture. Although obesity levels have been rising steadily over the last twenty years, many people do not actually see themselves as having a weight problem, despite being classified as “medically obese”. In the USA, 50% of the population are male. The probability of being male and obese is 8%. In total, 83% of the population are not obese, and 90% of people believe that they are not obese. Overall 95% of people are either obese, or believe that they are not obese, or both.

Find the probability that a (randomly chosen) US citizen is:

- (a) obese
- (b) female and obese
- (c) male ***and/or*** obese
- (d) obese ***and*** believes that they are *not obese*
- (e) obese ***and*** believes that they are obese.

Recipe for solving these kind of problems:

First formulate events:

let M = “Male” $F = M^c$ = “female”

let O = “is medically obese”

let B = “believe they are obese”

Next write down all the information given:

$$\begin{array}{ll} \mathbb{P}(M) = 0.5 & \mathbb{P}(O^c) = 0.83 \\ \mathbb{P}(M \cap O) = 0.08 & \mathbb{P}(O \cup B^c) = 0.95 \\ \mathbb{P}(B^c) = 0.9 & \mathbb{P}(B) = 1-0.9=0.1 \end{array}$$

Then write down what you want to know, and deduce how to get it from the information that you have.

Find the probability that a US citizen is:

(a) obese $\mathbb{P}(O) = 1 - \mathbb{P}(O^c) = 1 - 0.83 = 0.17$.

(b) female and obese **Want** $\mathbb{P}(F \cap O) = \mathbb{P}(M^c \cap O)$.

We know that $\mathbb{P}(M \cap O) + \mathbb{P}(M^c \cap O) = \mathbb{P}(O)$
i.e., $\mathbb{P}(M^c \cap O) + 0.08 = 0.17$
so, $\mathbb{P}(M^c \cap O) = 0.09$

(c) male **and/or** obese

$$\begin{aligned}\mathbb{P}(M \cup O) &= \mathbb{P}(M) + \mathbb{P}(O) - \mathbb{P}(M \cap O) \\ &= 0.50 + 0.17 - 0.08 \\ &= 0.59\end{aligned}$$

(d) obese **and** believes that they are not obese

Want: $\mathbb{P}(O \cap B^c)$
Know: $\mathbb{P}(O \cup B^c) = \mathbb{P}(O) + \mathbb{P}(B^c) - \mathbb{P}(O \cap B^c)$
 $0.95 = 0.17 + 0.90 - \mathbb{P}(O \cap B^c)$
 $\mathbb{P}(O \cap B^c) = 0.17 + 0.90 - 0.95$
 $= 0.12$

(e) obese **and** believes that they are obese

Want: $\mathbb{P}(O \cap B)$
Know: $\mathbb{P}(O \cap B^c) + \mathbb{P}(O \cap B) = \mathbb{P}(O)$
 $0.12 + \mathbb{P}(O \cap B) = 0.17$
 $\mathbb{P}(O \cap B) = 0.05$

3.3 Equally likely outcomes

Sometimes, all of the possible outcomes in a discrete finite sample space are equally likely. This makes it easy to calculate probabilities. If

i) $\Omega = \{\omega_1, \omega_2, \dots, \omega_k\}$

ii) $\mathbb{P}(\{\omega_1\}) = \mathbb{P}(\{\omega_2\}) = \dots = \mathbb{P}(\{\omega_k\}) = \frac{1}{k}$

iii) event $A \subset \Omega$ contains r possible outcomes

then $\mathbb{P}(A) = \frac{\text{number of outcomes in } A}{\text{number of outcomes in } \Omega} = \frac{r}{k}$.

We've already seen several examples of this. Here are a couple more.

Example 3.5. Toss a fair coin twice and observe the outcome each time. What is the probability of the event $T_1 = \{TH, TT\}$ that the first coin toss is a tail?

It is natural to assume that all 4 possible outcomes ($\Omega = \{HH, HT, TH, TT\}$) are equally likely, so

$$\mathbb{P}(\{HH\}) = \mathbb{P}(\{HT\}) = \mathbb{P}(\{TH\}) = \mathbb{P}(\{TT\}) = \frac{1}{4}$$

defines the probability measure on Ω that is appropriate for this experiment.

Event T_1 contains two of the four equally likely outcomes, so event T_1 occurs with probability $\frac{1}{2}$.

Example 3.6. For a 3-child family, possible outcomes for the gender of each child from oldest to youngest are:

$$\Omega = \{GGG, GGB, GBG, BGG, GBB, BGB, BBG, BBB\}$$

Real three-child families are very closely approximated by assuming that all outcomes in S are equally likely.

Let A be the event “oldest child is a girl” and let B be the event “youngest child is a boy”.

Then $A = \{GGG, GGB, GBG, GBB\}$,

$B = \{GGB, GBB, BGB, BBB\}$.

Event A contains 4 of the 8 equally likely outcomes, so event A occurs with probability $\mathbb{P}(A) = \frac{4}{8} = \frac{1}{2}$.

Similarly, event B has $\mathbb{P}(B) = \frac{1}{2}$.

The intersection $A \cap B = \{GGB, GBB\}$ has 2 of the 8 equally likely outcomes, so $\mathbb{P}(A \cap B) = \frac{2}{8} = \frac{1}{4}$.

3.3.1 Counting equally likely outcomes

To count the number of equally likely outcomes in an event, we often need to use **permutations** or **combinations**. If we are selecting r distinct objects from n distinct objects then:

- 1) when order matters (so the sequence (abc) is a different selection from (bac)):

$$\# \text{permutations} = {}^n P_r = n(n-1)(n-2) \dots (n-r+1) = \frac{n!}{(n-r)!}.$$

(n choices for first object, $(n-1)$ choices for second, etc.)

- 2) when order doesn't matter (so the sets $\{a, b, c\}$ and $\{b, a, c\}$ are the same):

$$\# \text{combinations} = {}^n C_r = \binom{n}{r} = \frac{{}^n P_r}{r!} = \frac{n!}{(n-r)!r!}.$$

(because ${}^n P_r$ counts all $r!$ possible orderings of the r objects).

Example 3.7. Bag of coloured balls

Four balls, each a different colour: Red, Blue, Green, Yellow

We want to choose two balls from the bag (without replacement) and observe the colour of each ball.

Possibilities: RB RG RY
 BR BG BY
 GR GB GY
 YR YB YG

If we assume that order is important, we can write 12 distinct sequences.. Easy to calculate this: there are 4 balls available for the first choice, and 3 available for the second choice: $4 \times 3 = 12$.

This relates to permutations:

$${}^n P_r = {}^4 P_2 = \frac{4!}{(4-2)!} = \frac{4 \times 3 \times 2 \times 1}{2 \times 1} = \frac{24}{2} = 12$$

If the order of the balls does not matter then there are only six outcomes, the six sets $\{R,B\}$, $\{R,G\}$, $\{R,Y\}$, $\{B,G\}$, $\{B,Y\}$, and $\{G,Y\}$.

This relates to combinations:

$${}^n C_r = {}^4 C_2 = \binom{4}{2} = \frac{4!}{(4-2)!2!} = \frac{4 \times 3 \times 2 \times 1}{(2 \times 1)(2 \times 1)} = \frac{24}{4} = 6$$

Example 3.8. If we draw five cards from a deck, what is the probability that none of the numbers on the cards match (numbered here 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13)?

Looking for $\mathbb{P}(\text{no cards have the same number}) = \mathbb{P}(A)$.

We will assume that the order in which the cards are chosen does not matter.

Total number of possible outcomes is:

$$52 \times 51 \times 50 \times 49 \times 48 / 5! \\ \text{or alternatively } {}^{52}C_5$$

Number of outcomes in A is number of ways of selecting 5 differently numbered cards:

$$\begin{aligned} & (\text{number of ways of selecting 5 differently numbered cards}) \\ &= 52 \times 48 \times 44 \times 40 \times 36 / 5! \\ & \text{or alternatively } {}^{13}C_5 \times 4^5 \end{aligned}$$

So

$$\mathbb{P}(A) = \frac{52 \times 48 \times 44 \times 40 \times 36}{52 \times 51 \times 50 \times 49 \times 48} = 0.5071.$$

How your answers above change if the order in which the cards were chosen mattered?

3.4 Exercises

3.4.1 Common sources of caffeine in the diet are coffee, tea and cola drinks. Suppose that

55% of adults drink coffee
25% of adults drink tea
45% of adults drink cola

and also that

15% drink both coffee and tea
5% drink all three beverages
25% drink both coffee and cola
5% drink only tea

- (a) Draw a Venn diagram showing this information.
- (b) What percentage of adults drink only cola?
- (c) What percentage drink none of these beverages?

3.4.2 *Drosophila*, a fruit fly commonly used in genetic studies, have many different characteristics based on differences in their underlying genetic makeup. Suppose that the fruit flies in a lab at the University of Auckland have the following attributes:

45% have short wings
30% have long legs
35% have spikey bristles

and also that

15% have short wings and long legs
10% have all three characteristics
25% have short wings and spikey bristles
10% only have long legs

- (a) Draw a Venn diagram showing this information.
- (b) What percentage of fruit flies have only spikey bristles (and not the other two characteristics)?
- (c) What percentage of flies have none of these characteristics?

3.4.3 Use the following data to answer the questions below:

Age and marital status of US women (in thousands)

	Age			Total
	18 to 24	25 to 64	65 and over	
Married	3,046	48,116	7,767	58,929
Never married	9,289	9,252	768	19,309
Widowed	19	2,425	8,636	11,080
Divorced	260	8,916	1,091	10,267
Total	12,614	68,709	18,262	99,585

- (a) Calculate the probability that a married woman is between 25 and 64.
- (b) Calculate the probability that a married woman is between 18 and 24.
- (c) Compare these probabilities. Which is higher? Is this what you would expect?

3.4.4 Give a mathematical proof or counterexample to the following claim:

if two events A and B satisfy $\mathbb{P}(A) > .5$ and $\mathbb{P}(B) > .5$ then $A \cap B \neq \emptyset$.

3.4.5 Suppose that $\mathbb{P}(A) = \frac{3}{4}$ and $\mathbb{P}(B) = \frac{1}{3}$.

- (a) Find the smallest value that $\mathbb{P}(A \cap B)$ can take.
- (b) Find the largest value that $\mathbb{P}(A \cap B)$ can take.
- (c) Find the greatest lower bound for $\mathbb{P}(A \cup B)$.
- (d) Find the least upper bound for $\mathbb{P}(A \cup B)$.

4 Conditional probability, Bayes' theorem and independence

By the end of this chapter you should be able to:

- recognize, write in symbolic notation and calculate conditional probabilities, including using Bayes theorem
- use the multiplication rule and the Partition theorem
- understand the difference between independent and mutually exclusive events and calculate probabilities using their properties.

4.1 Conditional probability

Suppose A and B are two events on the same sample space. There will often be dependence between A and B : that is, if we know that B has occurred, this changes our assessment of the chance that A will occur.

Example 4.1. Roll a die once.

Let event A = “get a 6”

Let event B = “get an even number”

If the die is fair, then $\mathbb{P}(A) = \frac{1}{6}$ and $\mathbb{P}(B) = \frac{1}{2}$

However, if we *know* that B has occurred, intuitively we would say “given this new information, the probability that A occurs is $\frac{1}{3}$ ”. We might be tempted to write

$$\mathbb{P}(A \text{ occurs given that the outcome is one of those in } B) = \frac{1}{3}$$

but this is not allowed as the thing that we have tried to measure the probability of is not an event (i.e. it is not a set of possible outcomes in Ω).

So how do we formally make sense of such an intuitive statement? This is what conditional probability is all about.

Example 4.2. Probabilities from tables of counts.

The following are the numbers of deaths from ischaemic heart disease in NZ in 2016.

		Sex		
		Male	Female	Total
Age	< 45	37	6	43
	45 – 64	485	122	607
	65 – 74	504	204	708
	75 – 84	760	471	1231
	85+	874	1199	2073
Total		2660	2002	4662

Let event A = “victim is female”. Let event B = “victim is < 45”.

Suppose we choose a person at random from those in the table.

$$\mathbb{P}(A) = \frac{\# \text{ female victims}}{\text{total } \# \text{ victims}} = \frac{2002}{4662} = 0.4294$$

But, if we choose people only from those under 45 years old, then intuitively:

The probability that the victim is female given that victim is < 45 is:

$$\frac{\# \text{ female victims} < 45}{\text{total } \# \text{ victims} < 45} = \frac{6}{43} = 0.1395 \approx 0.14.$$

We will write this as $\mathbb{P}(A | B) = 0.14$. We have conditioned on event B .

Conditioning on event B means **restricting attention** to the set of outcomes in B (i.e. the set for which B is true).

Think of $\mathbb{P}(A | B)$ as:
the chance of observing one of the elements in A ,
from the set of elements in **B only**.

If we look more closely at our calculation we can see the formal definition of conditional probability. Recall that the event A = “victim is female” and the event B = “victim is < 45”.

$$\begin{aligned}
 \mathbb{P}(A|B) &= \frac{\# \text{ female victims} < 45}{\text{total } \# \text{ victims} < 45} \\
 &= \frac{\# \text{ of outcomes in } A \cap B}{\# \text{ of outcomes in } B} \\
 &= \frac{(\# \text{ of outcomes in } A \cap B)/(\# \text{ of outcomes in } \Omega)}{(\# \text{ of outcomes in } B)/(\# \text{ of outcomes in } \Omega)} \\
 &= \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.
 \end{aligned}$$

This is our definition of conditional probability:

Definition: Let A and B be two events. The **conditional probability that event A occurs, given that event B has occurred**, is written $\mathbb{P}(A|B)$, and is given by

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

Read $\mathbb{P}(A|B)$ as “probability of A , given B ”.

Note: $\mathbb{P}(A|B)$ and $\mathbb{P}(A \cap B)$ are usually very different things, as can be seen from the definition of $\mathbb{P}(A|B)$. For example, if $\mathbb{P}(B) \in (0, 1)$ then $\mathbb{P}(B \cap B) = \mathbb{P}(B) < 1$, but $\mathbb{P}(B|B) = 1$!

$\mathbb{P}(A|B)$ is the probability that **event A occurs given that B has occurred**.

$\mathbb{P}(A \cap B)$ is the probability that **both events A and B occur**.

4.1.1 The language of conditional probability

Whenever we are calculating a probability we need to think about the sample space (S), the set that we are picking from. We can think of conditional probabilities as changing the sample space.

- $\mathbb{P}(A)$ means the probability of picking a sample point in set A when you are picking out of all the sample points in the whole sample space S .
- $\mathbb{P}(A|B)$ means the probability of picking a sample point in set A when you are only picking out of the sample points in set B .
- $\mathbb{P}(A \cap B)$ means the probability of picking a sample point that is in both sets A and B when you are picking out of all the sample points in the whole sample space S .

Always ask yourself “who are we picking from”?

Example 4.3. Let Ω =people in this class; A =people enrolled in an Arts degree; B =people with blue eyes; F =people in their first year of study at UoA.

Write each of the following statements in probability notation:

1. The probability that a person has blue eyes.
 $\mathbb{P}(B)$
2. The probability that a person has a blue eyes and is in their first year of study at UoA.
 $\mathbb{P}(B \cap F)$
3. The probability that a blue eyed person is in their first year of study at UoA.
 $\mathbb{P}(F|B)$
4. The probability that a person in their first year of study at UoA has blue eyes.
 $\mathbb{P}(B|F)$
5. The probability that a person is enrolled in their first year of study at UoA towards an Arts degree.
 $\mathbb{P}(F \cap A)$
6. The probability that a person enrolled in an Arts degree is in their first year of study at UoA and has blue eyes.
 $\mathbb{P}(F \cap B|A)$
7. The probability that a person enrolled in their first year of study at UoA towards an Arts degree has blue eyes.
 $\mathbb{P}(B|F \cap A)$

4.1.2 The multiplication rule and Partition Theorem

For any events A and B , we have the following multiplication rule:

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

Proof. Immediate from the definition of conditional probability:

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

and

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

Partition theorem take 2: the Multiplication Rule gives us a new statement of the Partition Theorem.

Theorem 4.1. If B_1, \dots, B_m partition B , then for any event $A \subset B$,

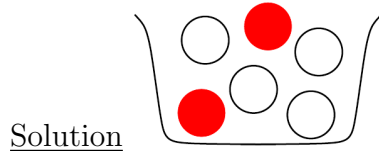
$$\mathbb{P}(A) = \sum_{i=1}^m \mathbb{P}(A \cap B_i) = \sum_{i=1}^m \mathbb{P}(A|B_i)\mathbb{P}(B_i)$$

Both formulations of the Partition Theorem are very widely used, but especially the conditional formulation $\sum_{i=1}^m \mathbb{P}(A|B_i)\mathbb{P}(B_i)$.

Note: Starting with a set of possible outcomes (sample space) S and a probability measure \mathbb{P} , when we condition on an event $B \subset S$ occurring we are saying that the outcome of our experiment is one of the outcomes in B . With this information, our new sample space (set of possible outcomes) is B . On this new sample space B , we also have a new probability measure, let's call it \mathbb{P}_B , that is a function from subsets of B to numbers in $[0, 1]$. This new probability measure \mathbb{P}_B is defined for $D \subset B$ by $\mathbb{P}_B(D) = \frac{\mathbb{P}(D)}{\mathbb{P}(B)}$ (as an exercise, check that defining \mathbb{P}_B in this way ensures that \mathbb{P}_B does satisfy the definition of a probability measure on B). Note that for any $A \subset S$, $A \cap B \subset B$ so by definition $\mathbb{P}_B(A \cap B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$ which is of course how we defined $\mathbb{P}(A|B)$. The notation $\mathbb{P}(A|B)$ is just a convenient and possibly more “readable” way of discussing conditional probability without referring to a change of sample space and probability measure every time we want to answer a question!

Example 4.4. Two balls are drawn at random without replacement from a box containing 4 white and 2 red balls. Find the probability that

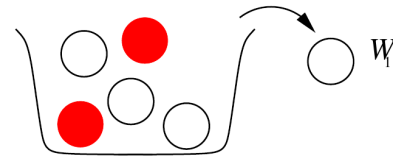
- (i) they are both white,
- (ii) the second ball is red.



Let event $W_i =$ “ i th ball is white”.

$$\text{i) } \mathbb{P}(W_1 \cap W_2) = \mathbb{P}(W_2 \cap W_1) = \mathbb{P}(W_2|W_1)\mathbb{P}(W_1)$$

$$\text{Now } \mathbb{P}(W_1) = \frac{4}{6} \quad \text{and} \quad \mathbb{P}(W_2|W_1) = \frac{3}{5}.$$



$$\text{So } \mathbb{P}(\text{both white}) = \mathbb{P}(W_1 \cap W_2) = \frac{3}{5} \times \frac{4}{6} = \frac{2}{5}.$$

Alternatively:

We can draw 2 white balls $\binom{4}{2} = 6$ ways.

We can we draw 2 balls $\binom{6}{2} = 15$ ways.

$$\text{Thus } \mathbb{P}(\text{both white}) = \frac{\binom{4}{2}}{\binom{6}{2}} = \frac{2}{5}.$$

or:

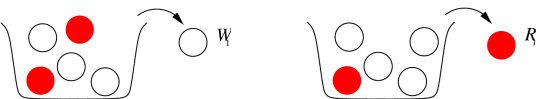
We can draw 2 white balls ${}^4P_2 = 12$ ways.

We can we draw 2 balls ${}^6P_2 = 30$ ways.

$$\text{Thus } \mathbb{P}(\text{both white}) = \frac{{}^4P_2}{{}^6P_2} = \frac{2}{5}.$$

ii) Event "2nd ball is red" is $R_2 = W_2^c = (W_1 \cap W_2^c) \cup (W_1^c \cap W_2^c)$.

$$\begin{aligned}\text{So, } \mathbb{P}(\text{2nd ball is red}) &= \mathbb{P}(W_1 \cap W_2^c) + \mathbb{P}(W_1^c \cap W_2^c) \quad (\text{mutually exclusive}) \\ &= \mathbb{P}(W_2^c|W_1)\mathbb{P}(W_1) + \mathbb{P}(W_2^c|W_1^c)\mathbb{P}(W_1^c)\end{aligned}$$



$$\begin{aligned}&= \frac{2}{5} \times \frac{4}{6} + \frac{1}{5} \times \frac{2}{6} \\ &= \frac{1}{3}\end{aligned}$$

Alternatively:

We can draw 2 red balls $2 \times 1 = 2$ ways.

We can draw a white then a red ball $4 \times 2 = 8$.

We can draw 2 balls $6 \times 5 = 30$ ways.

$$\text{Thus } \mathbb{P}(\text{second ball is red}) = \frac{2 + 8}{30} = \frac{1}{3}.$$

Note: Note that $\mathbb{P}(\text{1st ball is red})$ is also $\frac{1}{3}$. What is $\mathbb{P}(\text{ith ball is red})$, and why is it not necessary to do conditioning for this problem?

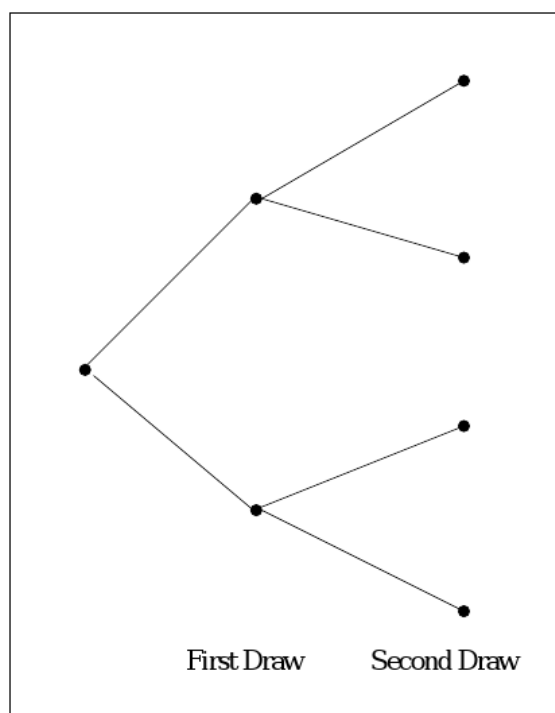
Label the balls $\{1, 2, 3, 4, 5, 6\}$ in any way you wish. Pick numbers from this set randomly without replacement. Any ordering is equally likely, so in particular the i th ball is equally likely to be any of the numbers from 1 to 6. Thus the probability that it is red is $1/3$.

4.1.3 Visualising conditional events

Probability trees are useful when [events happen in sequence](#).

Write conditional probabilities on the branches, and multiply to get probability of an intersection:

e.g. $\mathbb{P}(W_1 \cap W_2) = \frac{4}{6} \times \frac{3}{5}$.



4.2 Bayes' Theorem :reversing conditional probabilities

Remember the definition of condition probability: $\mathbb{P}(B|A) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)}$ provided $\mathbb{P}(A) > 0$.

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

Apply the multiplication rule to each side: $\mathbb{P}(B|A)\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B)$.

Thus,

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

This is the simplest form of Bayes' Theorem, named after Thomas Bayes (c. 1700), English clergyman and founder of Bayesian Statistics.



Figure 1: *According to Wikipedia this is the only portrait of Thomas Bayes, but even then there is some doubt as to whether it is actually him!*

Photo source: https://commons.wikimedia.org/wiki/File:Thomas_Bayes.gif

Bayes' Theorem allows us to “reverse” the conditioning, i.e., to express $\mathbb{P}(B|A)$ in terms of $\mathbb{P}(A|B)$.

This is very useful. For example, it might be easy to calculate

$$\mathbb{P}(\text{later event}|\text{earlier event}),$$

but we might only observe the later event and wish to deduce the probability that the earlier event occurred,

$$\mathbb{P}(\text{earlier event}|\text{later event}).$$

In the full version of Bayes' Theorem, we use the partition theorem to calculate $\mathbb{P}(A)$ in the denominator. A special case of this is to simply use B and B^c as the partition of S :

Example 4.5. Testing for HIV

Since the discovery of AIDS and HIV in the mid-1980s, blood tests for HIV have become a common medical procedure. For one particular HIV test (enzyme immunoassay), the probability that a person who is HIV positive actually tests positive is 0.997 (sensitivity), while the probability that a person who is HIV negative tests negative is 0.985 (specificity). In the US, HIV/AIDS prevalence is around 0.6% (2009 data).

Suppose a randomly selected individual in the US tests positive for HIV. What is the probability that they are actually HIV positive?

To solve this problem, follow the recipe:

Formulate events:

Let event P = “individual tests positive for HIV”

Let event HIV = “individual has HIV”

Write down all information given:

$$\begin{array}{ll} \mathbb{P}(P|HIV) = 0.997 & \text{so } \mathbb{P}(P^c|HIV) = 0.003 \\ \mathbb{P}(P^c|HIV^c) = 0.985 & \text{so } \mathbb{P}(P|HIV^c) = 0.015 \\ \mathbb{P}(HIV) = 0.006 & \text{so } \mathbb{P}(HIV^c) = 0.994 \end{array}$$

Write down what we’re looking for:

$$\mathbb{P}(HIV|P)$$

Compare this to what we know: $\mathbb{P}(P|HIV)$

We need to invert the conditioning, so use Bayes’ Theorem:

$$\begin{aligned} \mathbb{P}(HIV|P) &= \frac{\mathbb{P}(P|HIV)\mathbb{P}(HIV)}{\mathbb{P}(P)} \\ &= \frac{\mathbb{P}(P|HIV)\mathbb{P}(HIV)}{\mathbb{P}(P|HIV)\mathbb{P}(HIV) + \mathbb{P}(P|HIV^c)\mathbb{P}(HIV^c)} \\ &= \frac{0.997 \times 0.006}{0.997 \times 0.006 + 0.015 \times 0.994} \\ &= 0.286 \end{aligned}$$

What is the probability that someone who tests negative is actually HIV positive?

$$\begin{aligned}
 \mathbb{P}(HIV|P^c) &= \frac{\mathbb{P}(P^c|HIV)\mathbb{P}(HIV)}{\mathbb{P}(P^c)} \\
 &= \frac{\mathbb{P}(P^c|HIV)\mathbb{P}(HIV)}{\mathbb{P}(P^c|HIV)\mathbb{P}(HIV) + \mathbb{P}(P^c|HIV^c)\mathbb{P}(HIV^c)} \\
 &= \frac{0.003 \times 0.006}{0.003 \times 0.006 + 0.985 \times 0.994} \\
 &= 0.0000184
 \end{aligned}$$

Chou R, Huffman LH, Fu R, Smits AK, Korthuis PT (July 2005). *Screening for HIV: a review of the evidence for the U.S. Preventive Services Task Force*. Ann. Intern. Med. 143 (1): 55-73. PMID 15998755.

Here is the full statement of Bayes' Theorem for a general partition of Ω :

Theorem 4.2. Let B_1, B_2, \dots, B_m form a partition of Ω . Then for any event A , and any $j = 1, \dots, m$,

$$\mathbb{P}(B_j|A) = \frac{\mathbb{P}(A|B_j)\mathbb{P}(B_j)}{\sum_{i=1}^m \mathbb{P}(A|B_i)\mathbb{P}(B_i)}$$

Example 4.6. Study and work

The Graduate Destinations Report from Victoria University of Wellington reports on their Graduate destination survey. Two of the things the survey asks recent graduates about are their current study and employment. Many recent graduates are working part-time. Based on their 2012 survey (report available at www.victoria.ac.nz) it is estimated that 47% of those still in fulltime study are working part-time, 21% of those studying part-time are working part-time and 11% of those not studying are working part-time. Furthermore it is estimated that 22% of recent graduates are studying full-time and 8% are studying part-time.

Suppose you are chatting to a worker at McDonalds and discover that they are a recent graduate who is working part-time. What is the probability that they are studying full-time?

Formulate events:

Let:

FTS = “graduate in full-time study”

PTS = “graduate in part-time study”

N = “graduate not studying”

W = “graduate working part-time”

Write down all information given:

$$\mathbb{P}(W|FTS) = 0.47$$

$$\mathbb{P}(W|PTS) = 0.21$$

$$\mathbb{P}(W|N) = 0.11$$

$$\mathbb{P}(FTS) = 0.22$$

$$\mathbb{P}(PTS) = 0.08$$

Write down what we’re looking for: $\mathbb{P}(FTS|W)$

Compare this to what we know:

Need to invert the conditioning, so use Bayes’ Theorem:

$$\begin{aligned}\mathbb{P}(FTS|W) &= \frac{\mathbb{P}(W|FTS)\mathbb{P}(FTS)}{\mathbb{P}(W)} \\ &= \frac{\mathbb{P}(W|FTS)\mathbb{P}(FTS)}{\mathbb{P}(W|FTS)\mathbb{P}(FTS) + \mathbb{P}(W|PTS)\mathbb{P}(PTS) + \mathbb{P}(W|N)\mathbb{P}(N)} \\ &= \frac{0.47 \times 0.22}{0.47 \times 0.22 + 0.21 \times 0.08 + 0.11 \times (1 - 0.08 - 0.22)} \\ &= 0.5243\end{aligned}$$

So the part-time working recent graduate is studying full-time with probability 0.52.

What have we assumed here? Is it correct? Is it reasonable?

4.3 Independence

Suppose that $\mathbb{P}(A|B) = \mathbb{P}(A)$. This says that knowing that B occurs does not affect the probability of A occurring.

Furthermore if $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$ then $\mathbb{P}(B|A) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)} = \mathbb{P}(B)$

So whenever $\mathbb{P}(A|B) = \mathbb{P}(A)$, knowing whether A or B has occurred makes no difference to the probability of the other occurring. This is the intuitive notion that A and B are *independent* events.

Likewise we can show that if $\mathbb{P}(A|B) = \mathbb{P}(A)$ then $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

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

Definition: Events A and B are **independent** if

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

Note that making the definition in this way removes any concern about whether either event has zero probability.

In summary **events A and B are independent if**

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

Example 4.7. Roll a 6-sided die and observe the outcome.

Let $A = \{2, 4, 6\}$ be the event that the outcome is even, and

let $B = \{4, 5, 6\}$ be the event that the outcome is larger than 3.

Then $\mathbb{P}(A) = \mathbb{P}(B) = \frac{1}{2}$,

but $\mathbb{P}(A \cap B) = \mathbb{P}(\{4, 6\}) = \frac{1}{3} \neq \frac{1}{2} \times \frac{1}{2}$,

so **A and B are not independent.**

Example 4.8. Toss a fair coin and a fair 6-sided die and observe the outcome of each.

Sample space: $\Omega = \{H1, H2, H3, H4, H5, H6, T1, T2, T3, T4, T5, T6\}$, and most reasonable people would say that all of these outcomes are equally likely.

Let $A = \text{“coin toss is heads”}$ and $B = \text{“die roll is six”}$.

$$\text{Then } \mathbb{P}(A) = \mathbb{P}(\{H1, H2, H3, H4, H5, H6\}) = \frac{6}{12} = \frac{1}{2}$$

$$\text{and } \mathbb{P}(B) = \mathbb{P}(\{H6, T6\}) = \frac{2}{12} = \frac{1}{6}$$

$$\text{Now } \mathbb{P}(A \cap B) = \mathbb{P}(\{H6\}) = \frac{1}{12}. \quad \text{But } \mathbb{P}(A) \times \mathbb{P}(B) = \frac{1}{2} \times \frac{1}{6} = \frac{1}{12} \text{ also.}$$

So $\mathbb{P}(A \cap B) = \mathbb{P}(A) \times \mathbb{P}(B)$ and thus A and B are independent. This makes sense because the coin and die are physically independent.

Example 4.9. Job interviews

Suppose that jobs at a fast food chain are advertised in an Auckland newspaper. 70% of the applicants for the jobs have completed NCEA Level 1, and the probability that someone with this qualification will be hired is 0.4. The probability that an applicant has NCEA level 1 **OR** is hired for the job, is 0.82.

Let event $A = \text{“Applicant has NCEA Level 1”}$

Let event $B = \text{“Applicant is hired”}$

We know that:

$$\mathbb{P}(A) = 0.7, \quad \mathbb{P}(B|A) = 0.4, \quad \mathbb{P}(A \cup B) = 0.82$$

What is the probability that an applicant will get a job, regardless of their qualifications?

$$\begin{aligned} \mathbb{P}(A \cup B) &= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \\ &= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(B|A)\mathbb{P}(A) \end{aligned}$$

$$\begin{aligned} \mathbb{P}(B) &= \mathbb{P}(A \cup B) - \mathbb{P}(A) + \mathbb{P}(A \cap B) \\ &= \mathbb{P}(A \cup B) - \mathbb{P}(A) + \mathbb{P}(B|A)\mathbb{P}(A) \\ &= 0.82 - 0.7 + 0.4 \times 0.7 \\ &= 0.4 \end{aligned}$$

Are events A and B independent?

Yes: $\mathbb{P}(B) = \mathbb{P}(B|A)$, so $\mathbb{P}(A \cap B) = \mathbb{P}(B|A)\mathbb{P}(A) = \mathbb{P}(B)\mathbb{P}(A)$.

What does this mean?

According to this model the applicants' chances of getting a job are not affected by whether or not they have NCEA Level 1.

Notes:

- (1) If A and B are independent, so are: (i) A and B^c (ii) A^c and B (iii) A^c and B^c .

- (2) If A and B are mutually exclusive, they are usually **not** independent.

“Mutually exclusive” means $A \cap B = \emptyset$, which implies that $\mathbb{P}(A \cap B) = 0$.

“Independent” means $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

- (3) When events are *physically* independent, we assume that they are statistically/probabilistically independent as defined above.

For more than two events, we say:

Definition: Events A_1, A_2, \dots, A_n are **(mutually) independent** if

$$\mathbb{P}(A_1 \cap A_2 \cap \dots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2) \dots \mathbb{P}(A_n),$$

AND the same rule holds for every subcollection too.

e.g., events A_1, A_2, A_3 are (mutually) independent if

- (i) $\mathbb{P}(A_1 \cap A_2 \cap A_3) = \mathbb{P}(A_1)\mathbb{P}(A_2)\mathbb{P}(A_3)$, AND
- (ii) $\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1)\mathbb{P}(A_2)$,
 $\mathbb{P}(A_2 \cap A_3) = \mathbb{P}(A_2)\mathbb{P}(A_3)$,
 $\mathbb{P}(A_1 \cap A_3) = \mathbb{P}(A_1)\mathbb{P}(A_3)$.

In this course, when we roll a die more than once, or roll multiple dice etc., we assume that the outcomes of different rolls or different dice are mutually independent of each other.

Example 4.10. A jar contains 4 balls: one red, one white, one blue, and one multi-coloured red, white & blue ball. Draw one ball at random. Denote the sample points by the colour of the ball: r, w, b, m respectively (m for multi)

Let R = “ball has red on it”, W = “ball has white on it”, B = “ball has blue on it”.

$R = \{r, m\}$, so $\mathbb{P}(R) = \frac{2}{4} = \frac{1}{2}$. Similarly, $\mathbb{P}(W) = \mathbb{P}(B) = \frac{1}{2}$.

Now $\mathbb{P}(R \cap W) = \frac{1}{4}$.

But $\mathbb{P}(R) \times \mathbb{P}(W) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$, so $\mathbb{P}(R \cap W) = \mathbb{P}(R)\mathbb{P}(W)$.

Similarly, $\mathbb{P}(R \cap B) = \mathbb{P}(R)\mathbb{P}(B)$, and $\mathbb{P}(W \cap B) = \mathbb{P}(W)\mathbb{P}(B)$.

So R , W and B are pairwise independent.

But $\mathbb{P}(R \cap W \cap B) = \frac{1}{4}$

while $\mathbb{P}(R)\mathbb{P}(W)\mathbb{P}(B) = \frac{1}{8} \neq \mathbb{P}(R \cap W \cap B)$.

So R , W and B are *not* mutually independent, despite being pairwise independent.

4.4 Equally likely outcomes and independence

Example 4.11. Refer back to Example 3.6.

Recall that:

$$\Omega = \{GGG, GGB, GBG, BGG, GBB, BGB, BBG, BBB\},$$

$$A = \{GGG, GGB, GBG, GBB\}, \text{ and } B = \{GGB, GBB, BGB, BBB\}.$$

$$\text{Also } \mathbb{P}(A) = \frac{1}{2}, \mathbb{P}(B) = \frac{1}{2} \text{ and } \mathbb{P}(A \cap B) = \frac{1}{4}.$$

This means that $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$, so events A and B are **independent**.

Notes:

1. Suppose that

- the state space S consists of sequences (s_1, s_2, \dots, s_n) , with each coordinate s_i chosen from a finite set C_i of choices and no restrictions on possible combinations of coordinates; and
- all outcomes in S are equally likely.

Then:

- the events $A_1 = \{s_1 = a_1\}$, $A_2 = \{s_2 = a_2\}$, etc., are mutually independent; and
- if we write $C_1 = \{c_{1,1}, \dots, c_{1,m}\}$, then the events $B_{1,1} = \{s_1 = c_{1,1}\}, \dots, B_{1,m} = \{s_1 = c_{1,m}\}$ are equally likely, with $\mathbb{P}(B_{1,i}) = \frac{1}{|C_1|} = \frac{1}{m}$.

2. Similar statements hold for other coordinates, and more elaborate events such as

$$A = \{s_1 \in \{a_1, a_2\}\}, \quad B = \{s_2 = b \text{ and } s_4 \in \{d_1, d_2\}\}$$

are also independent.

4.5 Exercises

4.5.1 Use the following data to answer the questions below:

Age and marital status of US women (in thousands)

	Age			Total
	18 to 24	25 to 64	65 and over	
Married	3,046	48,116	7,767	58,929
Never married	9,289	9,252	768	19,309
Widowed	19	2,425	8,636	11,080
Divorced	260	8,916	1,091	10,267
Total	12,614	68,709	18,262	99,585

- (a) What is the probability that a randomly chosen women has never been married?
- (b) What is the probability that a woman over the age of 65 has never been married.?
- (c) Why do you think that these probabilities are different?

4.5.2 Pre-eclampsia is a serious condition which can occur during pregnancy. Women who have chronic hypertension (very high blood pressure) are at increased risk of developing the condition, as are women with high BMI (body mass index, a measure of a person's weight, relative to their height). Suppose that

- 8% of all pregnant women develop pre-eclampsia
- 20% of all pregnant women have high BMI
- 14% of all pregnant women have chronic hypertension

and also that

- 8% have chronic hypertension and high BMI
- 2% have chronic hypertension, high BMI and develop pre-eclampsia
- 3% have high BMI and develop pre-eclampsia
- 4% have chronic hypertension but not high BMI or pre-eclampsia

- (a) Let P be the event that a randomly chosen pregnant woman develops pre-eclampsia, B be the event that a randomly chosen pregnant woman has a high BMI and H be the event that a randomly chosen pregnant woman has chronic hypertension. Express all the percentages given as probabilities of these events (or combinations of them).
- (b) What percentage of pregnant women have high BMI, but not chronic hypertension or pre-eclampsia?(A Venn diagram may help)
- (c) What percentage of pregnant women don't have any of these three conditions?
- (d) What is the probability of a woman developing pre-eclampsia given that she has chronic hypertension?

- (e) Out of all women who have pre-eclampsia, what proportion don't have high BMI or chronic hypertension?
- (f) If a pregnant woman doesn't have high BMI or chronic hypertension, what is the probability she will develop pre-eclampsia?

4.5.3 After examining admissions data for graduate programmes (in law and engineering) at U. Auckland, a sociology student concluded that the University's admissions criteria were biased against women. His summary of the data is as follows:

A combined total of 200 students applied for these programmes, of which 100 were male and 100 female. Only 115 people were accepted, of which 55 were female. Let A be the event that a given student was accepted and F be the event that a given student is female.

- (a) Find $\mathbb{P}(A|F)$ and $\mathbb{P}(A|F^c)$.
- (b) Based on this data, do you agree with his conclusions about gender bias against women?

4.5.4 Further information about the data set from the previous question appears in an appendix to the sociology student's report. It contains the following additional information:

Out of the 200 students, 80 (=60 females + 20 males) applied to the law programme, while 120 (=80 males + 40 females) applied to the engineering programme.

Only 25 students were accepted into the law programme, of which 20 were female.

In engineering, 90 students were accepted, of which 35 were female.

Let L be the event that an applying student applied to the law programme (each student can apply to only one of the programmes).

- (a) Find $\mathbb{P}(A|F \cap L)$, $\mathbb{P}(A|F \cap L^c)$, $\mathbb{P}(A|F^c \cap L)$, $\mathbb{P}(A|F^c \cap L^c)$.
- (b) Based on this data, do you agree with the sociology student's conclusions that the admissions criteria were biased against females?

4.5.5 Recently a researcher at the University of Otago confirmed that the gene PTPN22 plays a role in the onset of rheumatoid arthritis. Suppose that 2% of all people have this gene, and that 25% of those with this gene will develop rheumatoid arthritis. Overall 1% of the population have rheumatoid arthritis.

- (a) Draw a decision tree to represent this situation.
- (b) What is the probability that a randomly selected individual from the population has the gene, but does not have rheumatoid arthritis?

- (c) What is the probability that a randomly selected individual from the population does not have the gene, but does have rheumatoid arthritis?
- (d) If an individual has rheumatoid arthritis, what is the probability that they also have the gene?

4.5.6 Suppose that in a remote village, 57% of the villagers are male, and 21% of the women have AIDS. In total, 17% of the villagers have the AIDS virus. Find the probability that a villager:

- (a) is female.
- (b) is female and has AIDS.
- (c) is male and has AIDS.
- (d) is male *and/or* has AIDS.
- (e) has AIDS, given that they are male.

4.5.7 A laboratory at the University of Otago studies a particular disorder which can be caused by a mutation (change in the DNA sequence) in a single gene. 1% of the population have this particular mutation, and 60% of those with the mutation develop the disorder. Overall, 0.8% of the population have the disorder.

- (a) Draw a decision tree to represent this situation.
- (b) What is the probability that a randomly selected individual from the population has the mutation, but does not have the disease?
- (c) What is the probability that a randomly selected individual from the population does not have the mutation, but does have the disease?
- (d) If an individual has the disease, what is the probability that they also have the mutation?

4.5.8 Suppose that in a population of cats, 5% have FIV (feline immunodeficiency virus). Tests for FIV are able to correctly identify 99% of positive cases, and 95% of negative cases.

- (a) Draw a decision tree to represent this situation.
- (b) What is the probability that a randomly selected cat will test positive for the disease?
- (c) What is the probability that a randomly selected cat which tests positive for FIV actually has the disease?
- (d) What is the probability that a randomly selected cat which tests negative for FIV actually has the disease?

- (e) Suppose that in (c) rather than considering a randomly selected cat, we want to know the probability that the next cat at the local veterinary clinic that tests positive for FIV actually has the disease. Would your answer change? Why or why not?

4.5.9 Give a mathematical proof or counterexample to the following claim:

If two events A and B satisfy $\mathbb{P}(A) > 0$, $\mathbb{P}(B) < 1$ and $A \subset B$ then A and B are dependent.

4.5.10 The distribution of blood types among males and females in the white / European / Pakeha population is approximately: 37% type A, 13% type B, 44% type O, and 6% type AB. Suppose that the blood types of people and their partners are independent.

- (a) Someone with blood type B can have transfusions of type B or type O blood. What is the probability that the partner of a person with type B blood can donate blood to them?
- (b) What is the probability that in a randomly chosen couple one person has type B blood and the other has type A?
- (c) What is the probability that in a randomly chosen couple the elder person has type A blood, and the younger person has type B?
- (d) What is the probability that at least one of a randomly chosen couple has type O blood?

4.5.11 The severity of colon cancer is rated on the Dukes scale, from A (least severe), to D (most severe). In the population of patients diagnosed with colon cancer, approximately 11% are Dukes A, 49% are Dukes B, 33% are Dukes C, and 7% are Dukes D at the time of surgery (when their primary tumour is removed).

- (a) What is the probability that a randomly selected colon cancer patient has a Dukes stage of A or B?
- (b) For Dukes B patients, the probability of relapse (i.e., the cancer coming back) is 16%. Overall the probability of relapse is, 31%. Is the probability of relapse independent of Dukes stage? Justify your answer.
- (c) The probability that a patient has chemotherapy is 23% (across all Dukes stages), but for Dukes C patients the probability of having chemotherapy is 53%.
 - (i) What is the probability that a randomly selected patient is Dukes C, but is not having chemotherapy?
 - (ii) If a patient is having chemotherapy, what is the probability that they are a Dukes C patient?

4.5.12 A laboratory has two strains of mouse, wildtype and mutant. The lab studies a cancer suppression gene which reduces the chance that a mouse with the gene will develop cancer. Suppose that 40% of the mice in the lab are mutants, and that $\frac{1}{6}$ of the wildtype mice have the gene. Let M be the event that “a randomly chosen mouse is from the mutant strain”, and let G be the event that “a randomly chosen mouse has the cancer suppressing gene”.

- (a) What is the probability that a randomly selected mouse is a wildtype, with the gene?
- (b) Suppose that out of all mice who have the gene, 75% are mutants.
 - (i) What is the probability that a randomly selected mouse has the gene?
 - (ii) Calculate the probability a randomly selected mouse is a mutant, with the gene.
 - (iii) Are events M and G independent? Justify your answer.

4.5.13 Suppose I deal two cards from a standard deck of 52 (assume that A=1, J=11, etc.)

- (a) What is the probability that the cards are a pair (i.e. they have the same number)?
- (b) What is the probability that the cards have both the same number, and the same colour?
- (c) What is the probability that the two cards have the same number *or* the same colour (or both)?
- (d) What is the probability that the numbers on the cards add to less than five?

4.5.14 In Roulette, the numbers 0 to 36 are coloured red (positive even numbers), black (odd numbers) or green (zero). Each time the wheel is spun, one of these colors will come up.

- (a) Suppose the Roulette wheel is spun twice and only the colour observed on each spin. Draw a Venn diagram to represent the sample space.
- (b) What is each colour’s probability of coming up when the wheel is spun?
- (c) Calculate the probability of each sample point in the sample space of part (a).
- (d) Let A be the event that at least one red comes up in two spins of the wheel, and let B be the event that at least one green comes up.
 - (i) What is the probability that both A and B occur?
 - (ii) What is the probability that either A or B occurs?
 - (iii) What is the probability that neither A nor B occurs?

4.5.15 A standard deck of cards contains 4 suits (Z =spades, D =diamonds, H =hearts, C =clubs), each consisting of 13 cards (Ace=1, King=13 etc.)

1. An experiment involves drawing 3 cards (without replacement) from the deck and observing *their suit only*, i.e. $\Omega = \{ZZZ, ZZD, \dots, CCC\}$.
 - (a) How many elements are there in the sample space Ω ?
 - (b) Are all the outcomes equally likely?
 - (c) Find the probability that all 3 cards are spades.
 - (d) Find the probability that all 3 cards are the same suit.
 - (e) Find the probability that all 3 cards are different suits.
 - (f) Find the following probabilities: $\mathbb{P}(\{ZZD\})$, $\mathbb{P}(\{ZDZ\})$, $\mathbb{P}(\{DZZ\})$, and describe in one sentence what $\mathbb{P}(\{ZZD\})$ means.
2. A different experiment involves drawing 3 cards (without replacement) from the deck and observing *their sum only*, i.e. $\Omega = \{3, \dots, 39\}$. Find $\mathbb{P}(\{39\})$ and $\mathbb{P}(\{4\})$.
3. A different experiment involves drawing 3 cards (without replacement) from the deck and observing *the number on each card*, i.e. $\Omega = \{(1, 1, 1), \dots, (13, 13, 13)\}$. Let A_n denote the event that the sum of the numbers on the 3 cards is n .
 - (a) Are all the outcomes equally likely?
 - (b) List the elements of A_1 , A_3 and A_4 in set notation (e.g. $A_{39} = \{(13, 13, 13)\}$).
 - (c) Do the events A_n , $n = 3, \dots, 39$ form a partition of Ω ? Explain why or why not.
 - (d) Find $\mathbb{P}(A_{39})$ and $\mathbb{P}(A_4)$.
4. A different experiment involves drawing 3 cards (without replacement) from the deck and observing *both the number and suit of each card*. Let A = “all 3 cards are spades”, B = “all 3 cards are the same suit”, C = “the sum of the numbers on the cards is 4”.
 - (a) How many elements are there in Ω ?
 - (b) Are all the outcomes equally likely?
 - (c) Let $x \in \Omega$. Find $\mathbb{P}(\{x\})$.
 - (d) Find $\mathbb{P}(A \cap B)$, $\mathbb{P}(A \cup B)$, $\mathbb{P}(B \cap C)$, $\mathbb{P}(B \cup C)$, and $\mathbb{P}(A \cup C)$.
 - (e) Identify which pairs of the events A , B , and C are independent.

4.5.16 A bin contains 6 white balls (numbered 1 to 6) and 3 red balls (numbered 7 to 9).

1. An experiment involves drawing 4 balls (without replacement) from the bin and observing *their colour only*, i.e. $\Omega = \{WWWW, WWWR, WWRW, \dots\}$.
 - (a) What does the outcome $WWWR$ mean in this context?
 - (b) How many elements are there in the sample space Ω ?
 - (c) Are all of these outcomes equally likely?
 - (d) Find the probability that all 4 balls are red.
 - (e) Find the probability that all 4 balls are white.
 - (f) Find the probability that all 4 balls are the same colour.
 - (g) Find the probability that you draw exactly 1 red ball.
 - (h) Find the probability that you draw at least one red and one white ball.
2. A different experiment involves drawing 4 balls (without replacement) from the bin and observing *only the sum of the numbers*.
 - (a) What is the sample space for this experiment?
 - (b) Find $\mathbb{P}(\{10\})$ and $\mathbb{P}(\{12\})$.
3. A different experiment involves drawing 4 balls (without replacement) from the bin and observing *the number and colour of each ball*. Let W_n denote the event that the colour of the n th ball drawn is white, and let A_n denote the event that the sum of the numbers on the 4 balls is n .
 - (a) How many elements are there in the sample space for this experiment?
 - (b) Find the probability of the event $W_1 \cap W_2 \cap W_3 \cap W_4$.
 - (c) Do the events W_n , $n = 1, \dots, 4$ form a partition of Ω ? Explain why or why not.
 - (d) Find $\mathbb{P}(A_{12})$ and $\mathbb{P}(A_{12} \cap W_1)$.

5 Discrete random variables, expectation and variance

By the end of this chapter you should be able to:

- distinguish between a probability mass function and a cumulative distribution function; know the properties of and be able to graph both types of functions
- be able to calculate the expected value, variance and standard deviation of a discrete random variable

5.1 Random variables

A random variable, X , assigns a real number to every possible outcome of a random experiment. To be precise a random variable is a function from Ω to \mathbb{R} (we write $X : \Omega \rightarrow \mathbb{R}$).

The random variable is discrete if the set of real values it can take is finite or countable, e.g. $\{0, 1, 2, \dots\}$. Any random variable defined on a discrete sample space Ω is discrete.

You are probably used to seeing notation of the form $\mathbb{P}(X = 0)$, however in this course we have said that a probability measure \mathbb{P} is a function from sets (subsets of Ω) to $[0, 1]$, so what does $\mathbb{P}(X = 0)$ mean?

Since X is a function from Ω to \mathbb{R} , the set $\{\omega \in \Omega : X(\omega) = 0\}$ is well defined and is a subset of S . Here's an example in terms of things that you should already be familiar with:

If $\Omega = \{-3, -2, -1, 0, 1, 2\}$ and $X(\omega) = \omega^2$, then $\{\omega \in \Omega : X(\omega) = 0\} = \{0\}$
 If $\Omega = \{-3, -2, -1, 0, 1, 2\}$ and $X(\omega) = \omega^2$, then $\{\omega \in \Omega : X(\omega) = 1\} = \{-1, 1\}$

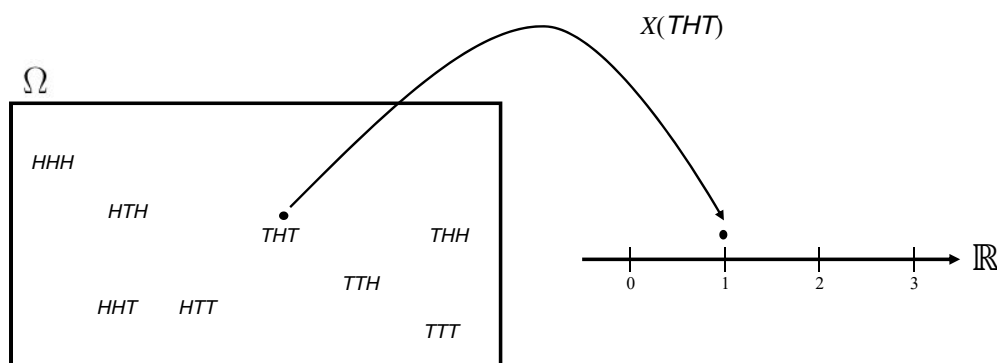
The notation $\mathbb{P}(X = 0)$ is shorthand for $\mathbb{P}(\{\omega : X(\omega) = 0\})$. More generally, if $E \subset \mathbb{R}$ then $\mathbb{P}(X \in E)$ is shorthand for $\mathbb{P}(\{\omega : X(\omega) \in E\})$. For example, $\mathbb{P}(X \leq 2) = \mathbb{P}(\{\omega : X(\omega) \leq 2\})$.

Example 5.1. An experiment involves tossing a coin 3 times and observing the result each time. There are $2^3 = 8$ possible outcomes in $\Omega = \{HHH, \dots, TTT\}$.

Let X be the number of heads observed, so $X(HHH) = 3$ and $X(HHT) = 2$ etc.

Then X is a random variable with possible values $\{0, 1, 2, 3\}$

and $\mathbb{P}(X = 0) = \mathbb{P}(\{TTT\}) = \frac{1}{8}$.



5.2 Probability mass function

Definition: The **probability mass function**, $f_X(x)$, for a discrete random variable X , is given by

$$f_X(x) = \mathbb{P}(X = x), \text{ for all possible outcomes } x \text{ of } X.$$

The probability mass function (p.m.f.) is sometimes called the *probability function*.

Example 5.2. Toss a fair coin once, and let X = number of heads. Then

$$\mathbb{P}(X = 0) = 0.5$$

$$\mathbb{P}(X = 1) = 0.5.$$

The p.m.f. of X is written

$$f_X(x) = \begin{cases} 0.5 & \text{if } x = 0 \\ 0.5 & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases} \quad \text{i.e. } f_X(0) = 0.5, \quad f_X(1) = 0.5, \quad f_X(y) = 0 \text{ if } y \neq 0, 1$$

By definition, the following properties must hold for any probability mass function $f_X(x)$:

i) $f_X(x) \geq 0$ for all x ; (probabilities are never negative)

ii) $\sum_x f_X(x) = 1$; (probabilities add to 1 overall)

iii) $\mathbb{P}(X \in A) = \sum_{x \in A} f_X(x)$.

Note: although any p.m.f. will satisfy all 3 of the above properties only i) and ii) are required for a function f , defined on a finite or countable set of real values, to be a p.m.f.

Example 5.3. A box contains six balls, two are labelled with the number 1, three are labelled with the number 2 and one is labelled with the number 4.
suppose we pick 1 ball from the box. Let X = the number on the ball. Then:

$$\mathbb{P}(X = 1) = \frac{1}{3} \qquad \mathbb{P}(X = 2) = \frac{1}{2} \qquad \mathbb{P}(X = 4) = \frac{1}{6}$$

The p.m.f. of X is written

$$f_X(x) = \begin{cases} \frac{1}{3} & \text{if } x = 1 \\ \frac{1}{2} & \text{if } x = 2 \\ \frac{1}{6} & \text{if } x = 4 \\ 0 & \text{otherwise} \end{cases} \quad \text{i.e. } f_X(1) = \frac{1}{3}, \quad f_X(2) = \frac{1}{2}, \quad f_X(4) = \frac{1}{6}, \quad f_X(y) = 0 \text{ if } y \neq 1, 2, 4$$

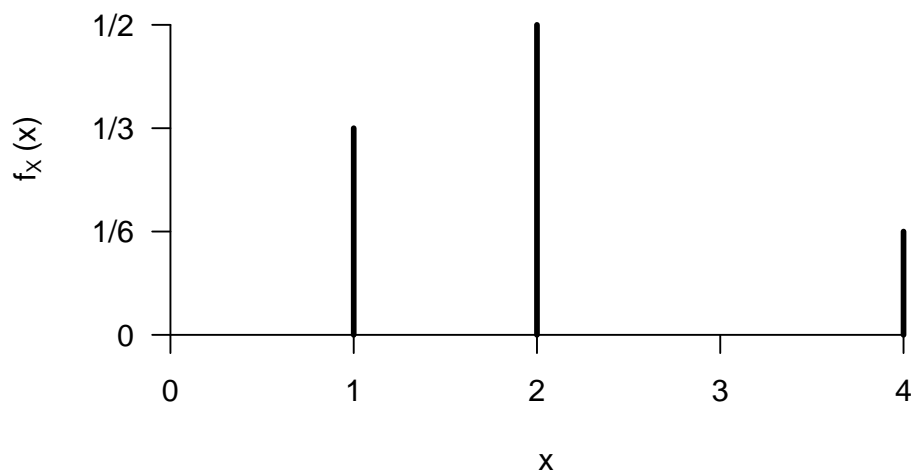
Check that this is a p.m.f.:

i) All probabilities are greater than zero

$$\text{ii) } \sum_x f_X(x) = \frac{1}{3} + \frac{1}{2} + \frac{1}{6} = 1$$

Note also that for example: $\mathbb{P}(X \in \{1, 4\}) = f_X(1) + f_X(4) = \frac{1}{3} + \frac{1}{6} = \frac{1}{2}$

A p.m.f. can be displayed on a bar chart:



5.3 The cumulative distribution function

We have defined the probability mass function, $f_X(x)$, as

$$f_X(x) = \mathbb{P}(X = x), \text{ for all possible outcomes } x \text{ of } X.$$

The cumulative distribution function, or just distribution function, written as $F_X(x)$, provides an alternative way of describing the distribution of X .

Definition: The cumulative distribution function (c.d.f.) is

$$F_X(x) = \mathbb{P}(X \leq x) \text{ for } -\infty < x < \infty$$

Either the cumulative distribution function, $F_X(x)$, or the probability mass function, $f_X(x)$, is sufficient to specify the distribution of X completely. (If we know one we can find the other)

Example 5.4. Continuing our example of the box of six balls with

$$f_X(x) = \begin{cases} \frac{1}{3} & \text{if } x = 1, \\ \frac{1}{2} & \text{if } x = 2, \\ \frac{1}{6} & \text{if } x = 4, \\ 0 & \text{otherwise.} \end{cases}$$

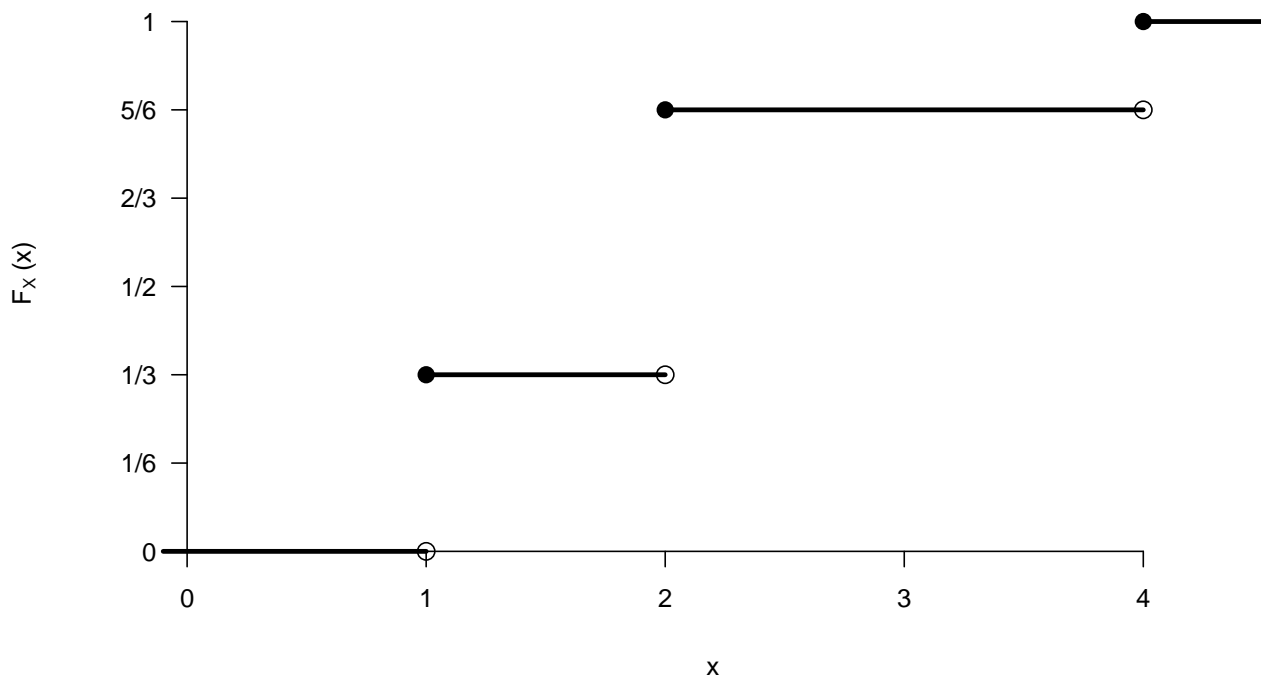
The c.d.f. is thus

$$F_X(x) = \mathbb{P}(X \leq x) = \begin{cases} 0 & \text{if } x < 1 \\ \frac{1}{3} & \text{if } 1 \leq x < 2 \\ \frac{1}{3} + \frac{1}{2} = \frac{5}{6} & \text{if } 2 \leq x < 4 \\ \frac{1}{3} + \frac{1}{2} + \frac{1}{6} = 1 & \text{if } x \geq 4 \end{cases}$$

$F_X(x)$ gives the cumulative probability up to and including point x .

So

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



Note that F_X is a **step function**:

it jumps by amount $f_X(y)$ at every point y for which $f_X(y) > 0$.

Note: As well as using the probability mass function to find the cumulative distribution function, we can also do the reverse:

$$\begin{aligned} f_X(x) &= \mathbb{P}(X = x) = \mathbb{P}(X \leq x) - \mathbb{P}(X \leq x - 1) \quad (\text{if } X \text{ takes integer values}) \\ &= F_X(x) - F_X(x - 1). \end{aligned}$$

More generally,

$$f_X(x) = F_X(x) - \lim_{u \uparrow x} F_X(u)$$

Example 5.5. Which of the following are probability mass functions? Justify your answer.

$$(a) \ f_X(x) = \begin{cases} \frac{x}{10} & \text{if } x = 1, 2, 3, 4 \\ 0 & \text{otherwise.} \end{cases}$$

$$(b) \ f_X(x) = \begin{cases} \frac{1}{x} & \text{if } x = 1, 2, 3, 4 \\ 0 & \text{otherwise.} \end{cases}$$

$$(c) \ f_X(x) = \begin{cases} 0.25 & \text{if } x = -1, -2, -3, -4 \\ 0 & \text{otherwise.} \end{cases}$$

$$(d) \ f_X(x) = \begin{cases} x & \text{if } x = -1, 0, 1 \\ 0 & \text{otherwise.} \end{cases}$$

For those that are probability mass functions find the cumulative distribution function.

Example 5.6. Which of the following are cumulative distribution functions? Justify your answer.

$$(a) \ F_X(x) = \begin{cases} 0 & x < -1 \\ 0.25 & -1 \leq x < 0 \\ 0.75 & x \geq 0 \end{cases}$$

$$(b) \ F_X(x) = \begin{cases} 0 & x < -1 \\ 0.5 & -1 \leq x < 0 \\ 1 & x \geq 0 \end{cases}$$

$$(c) \ F_X(x) = \begin{cases} 0 & x \leq 0 \\ 0.25 & 0 < x \leq 2 \\ 1 & x > 2 \end{cases}$$

$$(d) \ F_X(x) = \begin{cases} 0 & x < 1 \\ 0.25 & 1 \leq x < 3 \\ 1 & x \geq 3 \end{cases}$$

For those that are cumulative distribution functions find the probability mass function.

5.4 Expected value

Example 5.7. Suppose $X = \begin{cases} 1 & \text{with probability } 0.9, \\ -1 & \text{with probability } 0.1. \end{cases}$

X takes only the values 1 and -1 . What is the ‘average’ value of X ?

Using $\frac{1+(-1)}{2} = 0$ would not be useful, because

it ignores the fact that usually $X = 1$, and only occasionally is $X = -1$.

Instead, think of observing X many times, say 100 times.

Roughly 90 of these 100 times will have $X = 1$.

Roughly 10 of these 100 times will have $X = -1$.

Take the average of the 100 values: it will be roughly

$$\frac{90 \times 1 + 10 \times (-1)}{100},$$

i.e., $0.9 \times 1 + 0.1 \times (-1) = 0.8$.

This is why we take the average as

$$\mathbb{E}[X] = f_X(1) \times 1 + f_X(-1) \times (-1).$$

Definition: The **expected value**, or **mean**, of a discrete random variable X , can be written as $\mathbb{E}[X]$, or $E(X)$, or μ_X ,

and is given by

$$\mathbb{E}[X] = \sum_x x f_X(x) = \sum_x x \mathbb{P}(X = x)$$

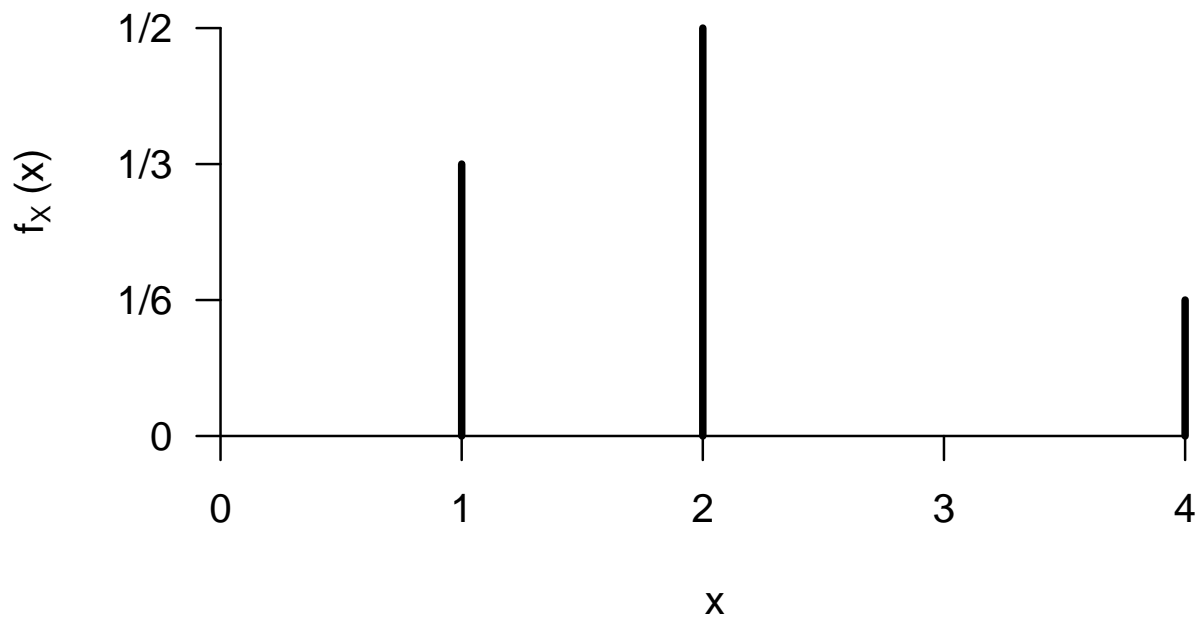
The expected value is a measure of the *centre*, or *average*, of the set of values that X can take, weighted according to the probability of each value.

$\mathbb{E}[X]$ is the average (mean) value we would get if we observed X many times.

Example 5.8. Consider the example of the box of six balls. Recall that the probability mass function is

$$f_X(x) = \begin{cases} \frac{1}{3} & \text{if } x = 1, \\ \frac{1}{2} & \text{if } x = 2, \\ \frac{1}{6} & \text{if } x = 4, \\ 0 & \text{otherwise.} \end{cases}$$

Calculate $\mathbb{E}[X]$ and mark this on the graph of the p.m.f..
What do you notice?



5.5 Expected value of a function of a random variable

Let X be a random variable, and let g be a function which transforms X .

Then $g(X)$ is also a random variable.

Example 5.9. $f_X(x) = \begin{cases} 0.75 & \text{if } x = 3 \\ 0.25 & \text{if } x = 8 \\ 0 & \text{otherwise} \end{cases}$

Let g be the square function, so that $g(x) = x^2$.

Then

$$f_X(g(x)) = \begin{cases} 0.75 & \text{if } x = 3^2 = 9 \\ 0.25 & \text{if } x = 8^2 = 64 \\ 0 & \text{otherwise} \end{cases}$$

So the average of $g(X)$ is:

$$\mathbb{E}[g(X)] = 0.75 \times 9 + 0.25 \times 64 = 22.75.$$

Theorem 5.1. For any function g , the expected value of $g(X)$ is given by

$$\mathbb{E}[g(X)] = \sum_x g(x)f_X(x) = \sum_x g(x)\mathbb{P}(X = x).$$

Theorem 5.2. Let a and b be constants, and let $g(x)$, $h(x)$ be functions. Then

i) $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$

ii) $\mathbb{E}[ag(X) + b] = a\mathbb{E}[g(X)] + b$

iii) $\mathbb{E}[ag(X) + bh(X)] = a\mathbb{E}[g(X)] + b\mathbb{E}[h(X)]$

Proofs:

$$\mathbb{E}[aX + b] = \sum_x (ax + b)f_X(x) \quad \text{Theorem 5.1}$$

$$= \sum_x (axf_X(x) + bf_X(x))$$

$$= \sum_x axf_X(x) + \sum_x bf_X(x)$$

$$= a \sum_x xf_X(x) + b \sum_x f_X(x)$$

$$= a\mathbb{E}[X] + b \quad \text{Definitions of Expected value and p.m.f}$$

Example 5.10. Let X be a random variable with the following probability function:

$$f_X(x) = \begin{cases} 0.6 & \text{if } x=1 \\ 0.3 & \text{if } x=2 \\ 0.1 & \text{if } x=3 \\ 0 & \text{otherwise} \end{cases}$$

Then

$$\mathbb{E}[X] = 1 \times 0.6 + 2 \times 0.3 + 3 \times 0.1 = 1.5$$

If $g(X) = 4X + 3$.

Then $\mathbb{E}[g(X)] = \mathbb{E}[4X + 3] = 4\mathbb{E}[X] + 3 = 9.0$

5.6 Expected value of sums and differences of random variables

Theorem 5.3. Let X and Y be random variables. Then

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$$

Example 5.11. Let X be the number obtained when you roll a fair six-sided die and Y be the number obtained when you roll a fair four-sided die.

What is the expected sum when rolling both die?

$$\mathbb{E}[X] = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = 3.5$$

$$\mathbb{E}[Y] = 1 \times \frac{1}{4} + 2 \times \frac{1}{4} + 3 \times \frac{1}{4} + 4 \times \frac{1}{4} = 2.5$$

$$\text{and thus } \mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y] = 3.5 + 2.5 = 6$$

More generally if X_1, X_2, \dots, X_n are random variables then

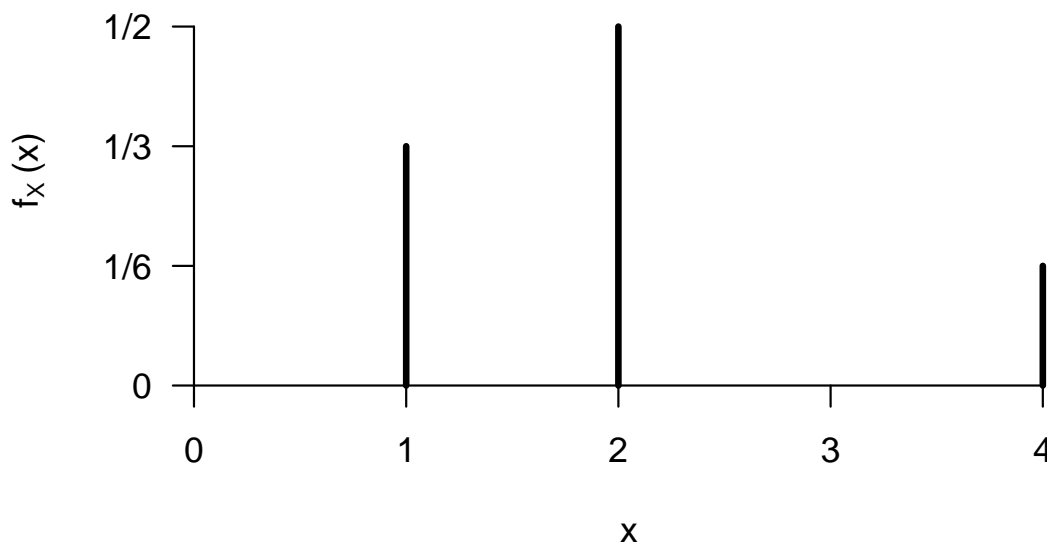
$$\mathbb{E}[X_1 + X_2 + \dots + X_n] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + \dots + \mathbb{E}[X_n].$$

5.7 Variance

The expected value of a p.m.f. provides information about the location of the centre of a probability distribution.

We are also interested in the spread of a distribution. This relates to how tightly packed the distribution is around the centre.

Consider our example of the box of six balls (again!) Previously we calculated that the expected value of X is 2.



From the bar plot we can see that this is a reasonable measure of the centre of the distribution, but it doesn't tell us anything about the spread that we see in the plot.

One way to measure the spread of a distribution is to calculate the *variance*.

Definition: The **variance** of a random variable X is written as either $\text{Var}(X)$ or σ_X^2 and is given by

$$\sigma_X^2 = \text{Var}(X) = \mathbb{E}[(X - \mu_X)^2] = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

Note: The variance is the square of the standard deviation of X ,

$$\text{sd}(X) = \sqrt{\text{Var}(X)} = \sqrt{\sigma_X^2} = \sigma_X.$$

The variance is a measure of how spread out the possible values of X are. It is the average squared distance between a value of X and the central (mean) value, μ_X .

$$\text{Var}(X) = \underbrace{\mathbb{E}}_{(2)} \underbrace{[(X - \mu_X)^2]}_{(1)}$$

- (1) Take distance from possible values of X to the central point, μ_X . Square it (*don't want positive and negative distances to "cancel"; want big deviations to be more costly than small ones*).
- (2) Then take the average over all values X can take: i.e., if we observed X many times, find what would be the average squared distance between X and μ_X .

Note: The mean, μ_X , and the variance, σ_X^2 , of X are just numbers: there is nothing random or variable about them.

For a discrete random variable,

$$\text{Var}(X) = \mathbb{E}[(X - \mu_X)^2] = \sum_x (x - \mu_X)^2 f_X(x) = \sum_x (x - \mu_X)^2 \mathbb{P}(X = x).$$

This uses the property of the expected value of a function of X :

$$\text{Var}(X) = \mathbb{E}[g(X)] \text{ where } g(X) = (X - \mu_X)^2.$$

Example 5.12. Let $f_X(x) = \begin{cases} 0.75 & \text{if } x = 3, \\ 0.25 & \text{if } x = 8. \end{cases}$

$$\begin{aligned} \mathbb{E}[X] &= \mu_X \\ &= 3 \times 0.75 + 8 \times 0.25 \\ &= 4.25 \end{aligned}$$

$$\begin{aligned} \text{Var}(X) &= \sigma_X^2 \\ &= 0.75 \times (3 - 4.25)^2 + 0.25 \times (8 - 4.25)^2 \\ &= 4.6875 \end{aligned}$$

When we observe X , we get either 3 or 8: this is random.

But μ_X is fixed at 4.25, and σ_X^2 is fixed at 4.6875, regardless of the outcome of X .

Theorem 5.4.

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \mathbb{E}[X^2] - \mu_X^2$$

Proof.

$$\begin{aligned}\text{Var}(X) &= \mathbb{E}[(X - \mu_X)^2] \quad \text{by definition} \\ &= \mathbb{E}\left[\underbrace{X^2}_{\text{r.v.}} - 2\underbrace{X}_{\text{r.v.}}\underbrace{\mu_X}_{\text{constant}} + \underbrace{\mu_X^2}_{\text{constant}}\right] \\ &= \mathbb{E}[X^2] - 2\mu_X\mathbb{E}[X] + \mu_X^2 \quad \text{by linearity of expectation} \\ &= \mathbb{E}[X^2] - 2\mu_X^2 + \mu_X^2 \\ &= \mathbb{E}[X^2] - \mu_X^2.\end{aligned}$$

Note: $\mathbb{E}[X^2] = \sum_x x^2 f_X(x) = \sum_x x^2 \mathbb{P}(X = x)$. This is not the same as $(\mathbb{E}[X])^2$:

$$\text{e.g., suppose that } f_X(x) = \begin{cases} 0.75 & \text{if } x = 3, \\ 0.25 & \text{if } x = 8. \end{cases}$$

then

$$\mu_X = \mathbb{E}[X] = 4.25$$

so

$$\mu_X^2 = (\mathbb{E}[X])^2 = (4.25)^2 = 18.0625$$

But

$$\mathbb{E}[X^2] = 3^2 \times 0.75 + 8^2 \times 0.25 = 22.75$$

Thus, $\mathbb{E}[X^2] \neq (\mathbb{E}[X])^2$ in general.

Example 5.13. Consider the following table which gives the investment return, or two different types of fund, under different economic conditions and the probability of each.

Economic Condition	Investment return		$\mathbb{P}(X = x)$	$\mathbb{P}(Y = y)$
	Passive fund X	Growth Fund Y		
Recession	-\$25	-\$200	0.2	0.2
Stable economy	\$50	\$60	0.5	0.5
Expanding economy	\$100	\$350	0.3	0.3

- (a) Sketch the probability mass function for both X and Y . How do you think the expected returns and variances will compare?
- (b) For each fund calculate the expected return ($\mathbb{E}[X]$ and $\mathbb{E}[Y]$), the variance of the return ($\text{Var}(X)$ and $\text{Var}(Y)$) and the standard deviation of the return ($\text{sd}(X)$ and $\text{sd}(Y)$).

(c) Which investment fund would you choose? Justify your answer.

Theorem 5.5. If a and b are constants, then

$$\text{Var}(aX + b) = a^2 \text{Var}(X).$$

Proof.

$$\begin{aligned} \text{Var}(aX + b) &= \mathbb{E} [(aX + b - \mathbb{E}[aX + b])^2] \\ &= \mathbb{E} [(aX + b - (a\mathbb{E}[X] + b))^2] \\ &= \mathbb{E} [(aX + b - a\mathbb{E}[X] - b)^2] \\ &= \mathbb{E} [(aX - a\mathbb{E}[X])^2] \\ &= \mathbb{E} [a^2(X - \mathbb{E}[X])^2] \\ &= a^2 \mathbb{E} [(X - \mathbb{E}[X])^2] \\ &= a^2 \text{Var}(X) \end{aligned}$$

Example 5.14. Let X be a random variable with the following probability function:

$$f_X(x) = \begin{cases} 0.1 & x = 1 \\ 0.6 & x = 8 \\ 0.3 & x = 27 \end{cases}$$

Suppose $g(x) = x^{\frac{1}{3}}$. Calculate the variance of $2g(X) + 3$.

$$\text{Let } Y = g(X). \text{ Then } f_Y(y) = \begin{cases} 0.1 & y = 1 \\ 0.6 & y = 2 \\ 0.3 & y = 3 \end{cases} \text{ and}$$

$$\mathbb{E}[g(X)] = \mathbb{E}[Y] = 0.1 \times 1 + 0.6 \times 2 + 0.3 \times 3 = 2.2$$

Note: $\mathbb{E}[X] = 0.1 \times 1 + 0.6 \times 8 + 0.3 \times 27 = 13$, so $\mathbb{E}[g(X)] \neq g(\mathbb{E}[X]) = (13)^{\frac{1}{3}}$.

$$\mathbb{E}[g(X)^2] = \mathbb{E}[Y^2] = 0.1 \times 1 + 0.6 \times 4 + 0.3 \times 9 = 5.2$$

$$\text{So } \text{Var}(g(X)) = \mathbb{E}[g(X)^2] - \mathbb{E}[g(X)]^2 = 5.2 - (2.2)^2 = 0.36.$$

$$\text{Finally, } \text{Var}(2g(X) + 3) = 2^2 \text{Var}(g(X)) = 4 \times 0.36 = 1.44.$$

5.8 Exercises

5.8.1 Given a parameter $\lambda \geq 0$, a very fishy discrete distribution has a probability function of the form

$$\mathbb{P}_\lambda(X = n) = \frac{e^{-\lambda}\lambda^n}{n!}, \quad n = 0, 1, 2, \dots$$

1. Verify that this really is a probability function.
2. For fixed n , and considering $\mathbb{P}_\lambda(X = n) = g_n(\lambda)$ as a function of $\lambda \geq 0$: For what values of λ is $g_n(\cdot)$ increasing and decreasing? Find the value of λ that maximises $\mathbb{P}_\lambda(X = 3)$.

5.8.2 An experiment involves rolling two fair dice. The first one is 4-sided and the second one is 6-sided, so there are 24 possible outcomes. Let X_1 denote the number showing on the first die, and Y denote the sum of the numbers showing on the two dice.

1. One of the possible outcomes for the experiment is $(1, 2)$. Find $X_1((1, 2))$ and $Y((1, 2))$.
2. Give the probability distribution for X_1 .
3. Find $\mathbb{E}[X_1]$ and $\text{Var}(X_1)$.
4. Find $\mathbb{E}[Y]$.

5.8.3 An ATM pin code consists of 4 numbers (from 0 to 9) in a specific sequence. A bank customer with a matching magnetic card enters this code to access their accounts. Suppose that your code consists of 4 different numbers and that immediately after you have used such a machine, a very smart thief is able to detect which numbers you entered, but not the order in which you entered them. The thief then steals your ATM card. Let X be the number of 4-digit codes that the thief has to try until being successful in accessing your account.

1. What are the possible values that X can take?
2. Find $\mathbb{E}[X]$.
3. In reality the machine stops the card from working if the wrong code is input 3 times in succession. Find $\mathbb{P}(X \leq 3)$.

5.8.4 An urn contains $n \geq 3$ balls of which k are red and $n - k$ are white. We draw 3 balls from the urn without replacement. Let X be the number of red balls chosen.

1. Find $\mathbb{P}(\{RRW\})$, $\mathbb{P}(\{RWR\})$ and $\mathbb{P}(\{WRR\})$. What does this suggest about the *order* in which balls are chosen?
2. Find $\mathbb{E}[X]$.

5.8.5 Let X be a random variable with the following probability function:

$$f_X(x) = \begin{cases} 0.3 & \text{if } x = 1, \\ 0.6 & \text{if } x = 4, \\ 0.1 & \text{if } x = 9, \\ 0 & \text{otherwise.} \end{cases}$$

- (i) Calculate the expected value of X .
- (ii) Calculate the variance of X .
- (iii) Suppose $g(X) = \sqrt{X}$. Calculate the mean (expected value) of $3g(X) + 2$.

5.8.6 A somewhat strange gambling game involves a six sided die, and a standard deck of 52 cards. A standard deck contains 13 cards (Ace, 1, 2,..., 10, Jack, Queen, King) of each suit (Clubs, Spades, Diamonds, Hearts). The Jack, Queen and King cards are known as “picture cards”. To play the game you roll the die, and then draw a card from the deck. If the number on the die matches the number of the card, you win \$3 (Aces count as ones). If the numbers don’t match, but the card is a picture card, you win \$1, unless you draw the Queen of Spades, in which case you win \$11. If the card is a non-picture card, and the numbers don’t match, you lose \$1.

- (a) What is the probability of drawing the Queen of Spades?
- (b) What is the probability of drawing a picture card which is *not* the Queen of Spades?
- (c) What is the probability that the number rolled on the die matches the number on the card that is drawn?
- (d) What is the expected value of this game?
- (e) Is this a game you would like to play? Please justify your answer.
- (f) How much would you have to win when the numbers match on the die and the card to make this a fair game? **Note:** a fair game is one where your expected return is 0.
- (g) Suppose that a Joker is added to the pack. The Joker automatically matches whatever number is on the die (i.e., you win \$3 if you draw the Joker).

- (i) What is the probability of winning at least \$1 when you play the game with the Joker in the pack?
- (ii) What is the expected value of the game now that the Joker has been added to the pack?
- (iii) Would you rather play the game with the Joker in the pack, or without it? Please justify your answer.

5.8.7 The popular casino game, Roulette, involves a spinning wheel with 37 slots, numbered 0,1,2,3,...,36. While the wheel is spinning, a small ball rolls into one of the slots. Gamblers place bets on what number they think the ball will be in when the wheel stops spinning. Suppose each bet costs \$1. If you win (i.e., if the ball ends up in the slot with your number), you receive \$35, otherwise you lose your bet (\$1).

- (i) If you bet \$1 each time, what is the expected value of this game?
- (ii) If you play the game 100 times, how much money would you expect to have won or lost?
- (iii) A “fair game” has an expected value of zero. How much would you have to win when your number comes up to make this a fair game?
- (iv) What is the probability that after 10 bets, you have won only twice?
- (v) What is the probability that you lose 20 times in a row?

5.8.8 A bin contains 6 white balls and 3 red balls. If you draw 4 balls from the bin without replacement, what is the expected number of white balls? What is the expected number of red balls?

5.8.9 The distribution of a random variable X is given by the following probability function.

$$\mathbb{P}(X = x) = \begin{cases} \frac{1}{2}, & \text{if } x = 0 \\ \frac{1}{4}, & \text{if } x = 1 \\ \frac{1}{8}, & \text{if } x = 2 \text{ or } 3 \\ 0, & \text{otherwise.} \end{cases}$$

1. Calculate the mean and variance of X .
2. Give the cumulative distribution function $P(X \leq x)$, and draw a picture of it.
3. Find the expected value of 2^X .
4. Find the expected value of $6X + 2^X$.

5.8.10 Let X be a discrete random variable with the following probability function:

$$f_X(x) = \begin{cases} 0.2 & \text{if } x = 0 \\ 0.4 & \text{if } x = 1 \\ 0.3 & \text{if } x = 2 \\ 0.1 & \text{if } x = 3 \\ 0 & \text{otherwise} \end{cases}$$

- (a) Find the probability that X is equal to 1.
- (b) Find the probability that X is greater than 1.
- (c) Find the distribution function, $F_X(x)$, for X .
- (d) Calculate the expected value of X .
- (e) Calculate the variance of X .

5.8.11 Real numbers have the following transitivity property: if $a < b$ and $b < c$ then $a < c$. Suppose we have 3 6-sided dice with the following numbers on their faces:

D1: 1,1,4,4,7,7

D2: 2,2,5,5,5,5

D3: 3,3,3,3,6,6

Roll each die once and let X_i be the number on die i .

- (a) Find $\mathbb{P}(X_2 > X_1)$.
- (b) Find $\mathbb{P}(X_3 > X_2)$
- (c) Find $\mathbb{P}(X_1 > X_3)$
- (d) Which die is most likely to show the highest number?

5.8.12 A random variable X has a probability function given by

$$f_X(x) = \begin{cases} .5, & \text{if } x = -1 \\ p, & \text{if } x = 0 \\ .2, & \text{if } x = 1 \\ 0, & \text{otherwise} \end{cases}$$

- (a) Find the value of p .

- (b) Find the expected value, $\mathbb{E}[X]$, of X .
- (c) Find the variance, $\text{Var}(X)$, of X .

5.8.13 On a game show called “Steal or no steal”, contestants can win up to \$200,000 as follows: There are 26 briefcases containing the following amounts (in dollars):

.10	1	2	5	10	20	50	75	100	200	300	500	750	1000	1500
2000	2500	3000	4000	5000	7500	10000	20000	50000	100000	200000.				

A contestant chooses one briefcase and holds it as their own. This briefcase remains unopened. The contestant then chooses 5 more cases which will be opened, revealing the amounts of money in those cases and removing these amounts from the game. At this point the “banker” (one of the organisers of the game) offers the contestant some money to buy the briefcase that the contestant is holding. The amount offered by the banker will depend on the amounts of money remaining in the game. If the contestant accepts that deal then they receive the offered amount and the game finishes. Otherwise the contestant continues to open more briefcases (removing the amounts of money inside them from the game) and the banker continues to make offers until either: the contestant accepts the offer made by the banker, or there are no remaining briefcases to open (at which point the contestant opens the briefcase they are holding and wins the amount inside - which is a known amount since all of the other briefcases have been opened!)

- (a) Suppose your strategy is to never accept the deal offered by the banker. What are your expected winnings from playing this game?
- (b) Suppose that the first 5 cases that are opened reveal the amounts .1, 1, 2, 100000, 200000, what is the expected amount in the briefcase that you are holding?
- (c) The banker typically makes an offer that is less than the expected amount in the briefcase that you are holding. Why do you think this is the case? (In considering this question you may like to think about what offer from the banker YOU would accept if you get to the point where the only two amounts remaining are .1 and 200000.)

6 Discrete distributions

By the end of this chapter you should be able to:

- list the key properties of,
- recognize when it is appropriate to model a situation using, and
- solve problems involving probabilities, mean and variance of

the following discrete distributions:

- Bernoulli
- Binomial
- Geometric
- Negative Binomial
- Hypergeometric
- Discrete Uniform
- Poisson

By definition, a random variable X is a function defined on a sample space S . Often, however, we are interested only in the properties of X that can be measured using probabilities, such as the expected value $\mathbb{E}(X)$, the variance $\text{Var}(X)$, or the probability mass function $f_X(x)$. These properties are determined by the *distribution* of X .

Two discrete random variables X, X' have the same distribution if $\mathbb{P}(X = x) = \mathbb{P}(X' = x)$ for every x – in other words, if they have the same p.m.f., or equivalently the same c.d.f. However, there is no need for X and X' to be equal. In fact, X and X' do not even need to arise from the same random experiment, and they could be defined on two different sample spaces. It is often interesting to see when two apparently different random variables turn out to have the same distribution.

In this chapter, we introduce several of the most common discrete distributions. Random variables having these distributions occur naturally in many different random experiments. However they arise, all such random variables must share the properties encoded in their distribution.

6.1 Bernoulli random variables

Definition: A random experiment is called a **set of Bernoulli trials** if it consists of several trials such that:

- i) Each trial has only 2 possible outcomes (usually called “Success” and “Failure”);
- ii) The probability of success, p , remains constant for all trials;
- iii) The trials are independent.

Example 6.1. Repeated tossing of a fair coin: success = “head”, failure = “tail”.

Each toss is a Bernoulli trial with $\mathbb{P}(\text{success on } i\text{th trial}) = \frac{1}{2}$.

Example 6.2. Repeated tossing of a fair die: success = “6”, failure= “not 6”.

Each toss is a Bernoulli trial with $\mathbb{P}(\text{success on } i\text{th trial}) = \frac{1}{6}$.

Definition: A random variable Y is called a **Bernoulli random variable** if it has only two possible values, 0 and 1.

$Y \sim \text{Bernoulli}(p)$ means $\mathbb{P}(Y = 1) = p$ and $\mathbb{P}(Y = 0) = 1 - p$.

and p is the **parameter** of the distribution.

Definition: An **indicator variable** (Π_A) for the event A is a random variable which takes the value 1 when event A occurs and 0 otherwise.

Thus in the examples above, each of the outcomes can be coded using 1 to represent success, and 0 to represent failure.

For **Example 6.1** the probability of each occurrence is given by:

$$\begin{aligned}\mathbb{P}(Y = 1) &= \mathbb{P}(\text{“success”}) = \mathbb{P}(\text{“head”}) = \frac{1}{2} \text{ and} \\ \mathbb{P}(Y = 0) &= \mathbb{P}(\text{“failure”}) = \mathbb{P}(\text{“tail”}) = 1 - \frac{1}{2} = \frac{1}{2}.\end{aligned}$$

For **Example 6.2** the probability of each occurrence is given by:

$$\begin{aligned}\mathbb{P}(Y = 1) &= \mathbb{P}(\text{“success”}) = \mathbb{P}(\text{“toss a 6”}) = \frac{1}{6} \text{ and} \\ \mathbb{P}(Y = 0) &= \mathbb{P}(\text{“failure”}) = \mathbb{P}(\text{“not a 6”}) = 1 - \frac{1}{6} = \frac{5}{6}.\end{aligned}$$

6.1.1 Bernoulli probability mass function

A random variable $Y \sim \text{Bernoulli}(p)$ has a probability function given by:

$$f_Y(y) = \begin{cases} p & \text{if } y = 1, \\ 1 - p & \text{if } y = 0. \end{cases}$$

6.1.2 Bernoulli cumulative distribution function

A random variable $Y \sim \text{Bernoulli}(p)$ has a cumulative distribution function given by:

$$F_Y(y) = \begin{cases} 0 & \text{if } y < 0, \\ 1 - p & \text{if } 0 \leq y < 1, \\ 1 & \text{if } y \geq 1. \end{cases}$$

6.1.3 Mean and variance of a Bernoulli random variable

If $Y \sim \text{Bernoulli}(p)$ then:

$$\mathbb{E}[Y] = 0 \times (1 - p) + 1 \times p = p$$

and:

$$\text{Var}(Y) = \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2 = 0^2 \times (1 - p) + 1^2 \times p - p^2 = p - p^2 = p(1 - p)$$

Example 6.3. Suppose that 20% of students have a student loan larger than \$10,000.

A random experiment involves randomly choosing students and finding out if they have a student loan larger than \$10,000.

Is this a sequence of Bernoulli trials?

6.2 Binomial distribution

Example 6.4. Suppose that 20% of students have a student loan larger than \$10,000. If we choose six students at random, what is the probability that exactly four of them have student loans of more than \$10,000?

Experiment 1: For all of the 6 students observe whether or not their student loan is greater than \$10,000.

Let A_i = “student i has loan greater than \$10,000”, $i = 1, 2, \dots, 6$.

We are given the information that $\mathbb{P}(A_i) = 0.2$ for each A_i , so the events A_i are equally likely and independent.

Let E be the event that exactly four of the students have a student loan greater than \$10,000

i.e. four students have loans greater than \$10,000, and two students have loans less than or equal to \$10,000.

How many outcomes are in E ? ${}^6C_4 = \frac{6!}{(6-4)!4!} = \frac{720}{48} = 15$

What is the probability of each outcome in E ?

$$0.2 \times 0.2 \times 0.2 \times 0.2 \times 0.8 \times 0.8 = (0.2)^4 \times (0.8)^2 = 0.001024$$

So the probability of getting exactly four (out of six) students with loans greater than \$10,000 is:

$$\mathbb{P}(E) = {}^6C_4 \times (0.2)^4 \times (0.8)^2$$

Experiment 2: We observe only the number of students that have a student loan greater than \$10,000. $S = \{0, 1, 2, 3, 4, 5, 6\}$.

What is the probability that the outcome is 4?

$$\binom{6}{4}(0.2)^4(0.8)^2$$

This problem is an example of the Binomial distribution.

The Binomial distribution describes the behaviour of a random variable which records the number of successes in a fixed number of (independent and identical) trials.

Here we are interested in the number of students with loans greater than \$10,000 (“successes”), out of the six students (trials). The outcome of each trial was independent of all other trials, and all had the same probability of success.

Definition: Let X be the number of successes in n independent Bernoulli trials each with probability of success $= p$.

Then X has the Binomial distribution with parameters n and p .

We write $X \sim \text{Binomial}(n, p)$, or $X \sim \text{Bin}(n, p)$.

Thus the Binomial distribution counts the number of successes in a fixed number of Bernoulli trials

6.2.1 Binomial probability mass function

A random variable $X \sim \text{Binomial}(n, p)$ has a p.m.f. given by:

$$f_X(x) = \mathbb{P}(X = x) = \binom{n}{x} p^x (1 - p)^{n-x} \quad \text{for } x = 0, 1, \dots, n.$$

Explanation:

What is the probability of success on each trial? p

What is the probability of failure on each trial? $1 - p$

If we have x successes out of n trials how many failures are there? $n - x$

This means that each of the outcomes with x successes and $n - x$ failures has probability

$$p^x (1 - p)^{n-x}$$

There are $\binom{n}{x}$ possible outcomes with x successes and $(n - x)$ failures because we must select x trials to be our “successes”, out of n trials in total.

What is probability of getting x successes out of n trials?

$$\begin{aligned} \mathbb{P}(\# \text{successes} = x) &= (\# \text{outcomes with } x \text{ successes}) \times (\text{prob. of each such outcome}) \\ &= \binom{n}{x} p^x (1 - p)^{n-x} \end{aligned}$$

Example 6.5. Student loans example revisited

X = number of students with student loans > \$10,000.

n = number of trials = 6

p = probability of success = 0.2

So $X \sim \text{Binomial}(n = 6, p = 0.2)$

The p.m.f. is:

$$f_X(x) = \mathbb{P}(X = x) = \binom{6}{x} 0.2^x (1 - 0.2)^{6-x} \quad \text{for } x = 0, 1, \dots, 6.$$

What is the probability of getting (exactly) five students with student loans greater than \$10,000 out of a total of six students?

$$\mathbb{P}(X = 5) = \binom{6}{5} (0.2)^5 (1 - 0.2)^{6-5} = 0.001536$$

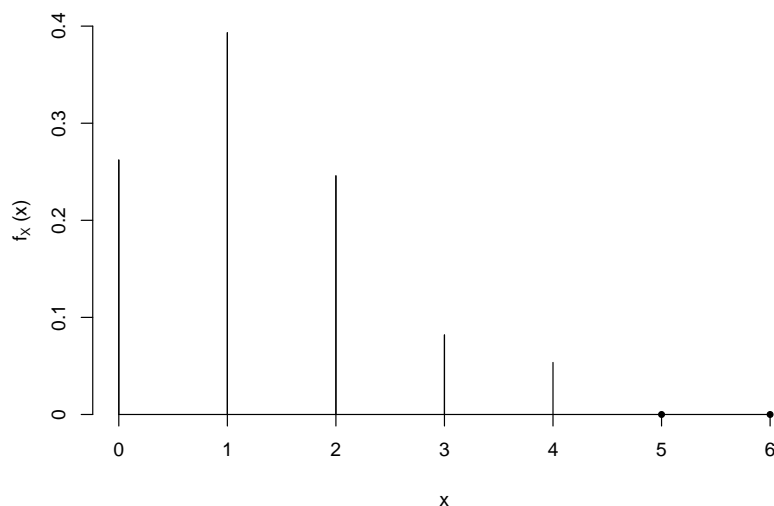
What is the probability of getting no students with student loans greater than \$10,000 out of a total of six students?

$$\mathbb{P}(X = 0) = \binom{6}{0} (0.2)^0 (1 - 0.2)^{6-0} = 0.2621$$

What is the probability of getting at least one student with a student loan greater than \$10,000 out of a total of six students?

$$\mathbb{P}(X \geq 1) = 1 - \mathbb{P}(X = 0) = 1 - 0.2621 = 0.7329$$

We can also plot the p.m.f.



6.2.2 Binomial cumulative distribution function

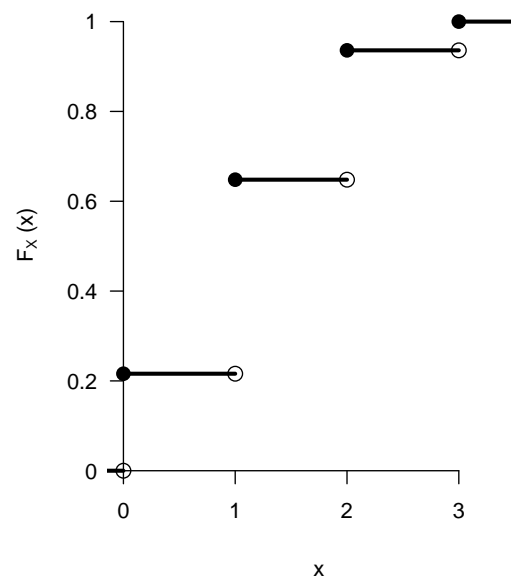
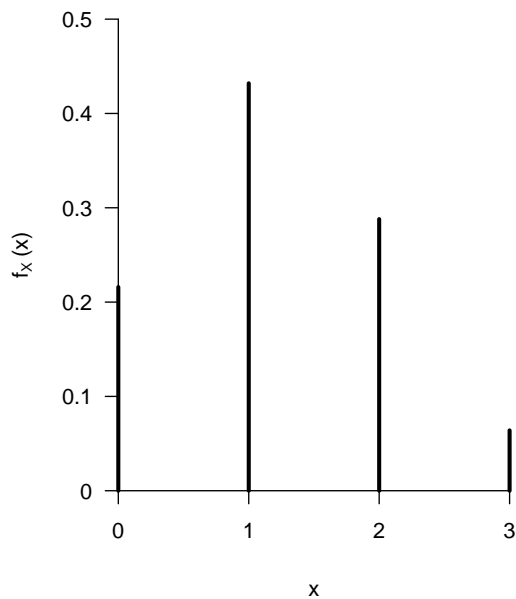
Example 6.6. Let $X \sim \text{Binomial}(3, 0.4)$.

The p.m.f. is $f_X(x) = \mathbb{P}(X = x) = \binom{3}{x} 0.4^x (0.6)^{3-x}$ for $x = 0, 1, \dots, 3$.

x	0	1	2	3
$f_X(x) = \mathbb{P}(X = x)$	0.216	0.432	0.288	0.064

Recall that $F_X(x) = \sum_{y \leq x} f_X(y)$

$$\text{So } F_X(x) = \mathbb{P}(X \leq x) = \begin{cases} 0 & \text{if } x < 0, \\ 0.216 & \text{if } 0 \leq x < 1, \\ 0.216 + 0.432 = 0.648 & \text{if } 1 \leq x < 2, \\ 0.216 + 0.432 + 0.288 = 0.936 & \text{if } 2 \leq x < 3, \\ 0.216 + 0.432 + 0.288 + 0.064 = 1 & \text{if } x \geq 3. \end{cases}$$



6.2.3 Mean and variance of the Binomial distribution

Example 6.7. Binomial distribution, $n = 3$, $p = 0.4$

$$\begin{aligned}\mathbb{E}[X] &= 0 \times 0.216 + 1 \times 0.432 + 2 \times 0.288 + 3 \times 0.064 \\ &= 1.2\end{aligned}$$

$$\begin{aligned}\mathbb{E}[X^2] &= 0^2 \times 0.216 + 1^2 \times 0.432 + 2^2 \times 0.288 + 3^2 \times 0.064 \\ &= 2.16\end{aligned}$$

$$\begin{aligned}\text{Var}(X) &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\ &= 2.16 - 1.2^2 \\ &= 0.72\end{aligned}$$

There are also shortcuts to calculate these values for the Binomial, and other distributions.

If $X \sim \text{Binomial}(n, p)$, then

$$\mathbb{E}[X] = np$$

$$\text{Var}(X) = np(1 - p) = npq, \text{ where } q = 1 - p$$

Proof:

If $X \sim \text{Binomial}(n, p)$ then we can think of X as the sum of n Bernoulli random variables. That is $X = Y_1 + Y_2 + \dots + Y_n$ where each $Y_i \sim \text{Bernoulli}(p)$.

Recall that a success in a Bernoulli trial means that $Y_i = 1$ this mean that $Y_1 + Y_2 + \dots + Y_n$ will be the same as the total number of successes.

From Section 6.1.3 we know that if $Y_i \sim \text{Bernoulli}(p)$ then $\mathbb{E}[Y_i] = p$.

Thus

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}[Y_1 + Y_2 + \dots + Y_n] \\ &= \mathbb{E}[Y_1] + \mathbb{E}[Y_2] + \dots + \mathbb{E}[Y_n] && \text{from Theorem 5.3} \\ &= p + p + \dots + p \\ &= np\end{aligned}$$

The proof of the variance formula is similar, but since it uses a formula from Chapter 8, it will be postponed until Example 8.5.

A more mathematical proof that $\mathbb{E}[X] = np$. (*non-examinable*)

$$\begin{aligned}
\mathbb{E}[X] &= \sum_x x f_X(x) = \sum_{x=0}^n x \binom{n}{x} p^x (1-p)^{n-x} \text{ (by definition)} \\
&= \sum_{x=1}^n x \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \text{ (taking out the first term since it is 0)} \\
&= \sum_{x=1}^n \frac{n!}{(x-1)!(n-x)!} p^x (1-p)^{n-x} \text{ (cancelling)} \\
&= np \sum_{x=1}^n \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1} (1-p)^{n-x} \text{ (taking out a common factor of } np) \\
&= np \sum_{y=0}^{n-1} \frac{(n-1)!}{y!(n-y-1)!} p^y (1-p)^{n-y-1}
\end{aligned}$$

To get the last expression we set $y = x - 1$. So since the summation runs from $x = 1$ to $x = n$, it becomes the summation from $y = 0$ to $y = n - 1$. Within the summation $x - 1$ is written as y , and $n - x$ becomes $n - (y + 1) = n - y - 1$.

Finally we observe that if Y is a random variable with the Binomial $(n - 1, p)$ distribution, then it has a probability mass function

$$f_Y(y) = \begin{cases} \frac{(n-1)!}{y!(n-y-1)!} p^y (1-p)^{n-y-1}, & y = 0, 1, 2, \dots, n-1, \\ 0 & \text{otherwise.} \end{cases}$$

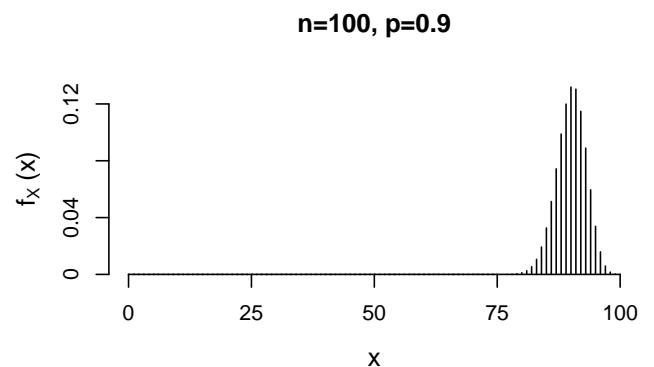
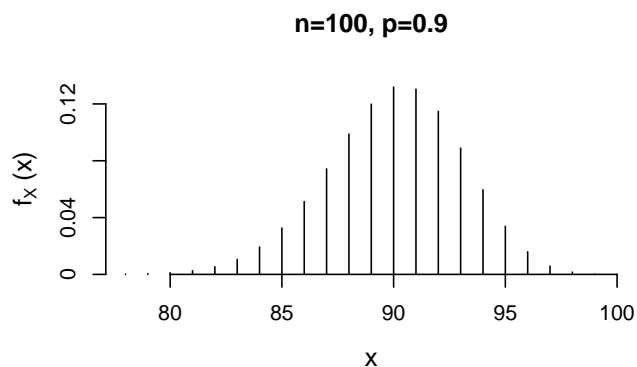
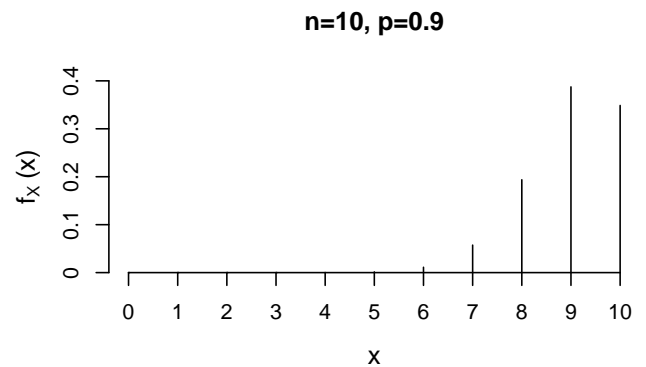
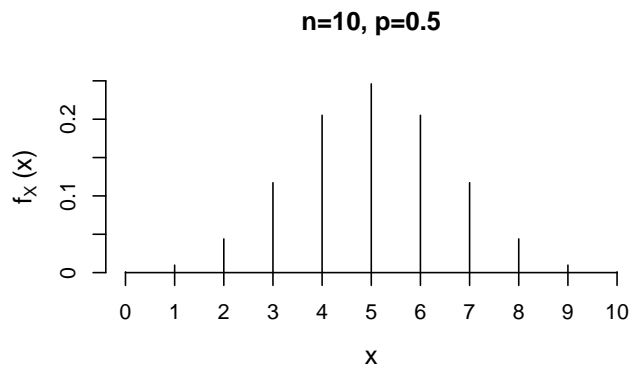
and by definition of the probability mass function $\sum_y f_Y(y) = 1$ so that

$$\sum_{y=0}^{n-1} \frac{(n-1)!}{y!(n-y-1)!} p^y (1-p)^{n-y-1} = 1$$

and hence $\mathbb{E}[X] = np$.

6.2.4 Shape of the Binomial distribution

The shape of the Binomial distribution depends upon the values of n and p . For small n , the distribution is almost symmetrical for values of p close to 0.5, but highly skewed for values of p close to 0 or 1. As n increases, the distribution becomes more and more symmetrical, and there is noticeable skew only if p is very close to 0 or 1. Plots of the probability functions for various values of n and p are shown below.



6.3 Geometric distribution

- If I have a fair 6-sided die what is the probability that I “roll a 6”? $\frac{1}{6}$
- Suppose I roll the die 4 times. Let X be the number of ‘6’s.
What is the distribution of X ? $X \sim \text{Binomial}(n = 4, p = \frac{1}{6})$
- Many board games require you to “roll a 6” before you can start the game. Suppose that I let X be the number of rolls before the one on which I get a ‘6’
 - Is this still a sequence of Bernoulli trials? **Yes**
 - Does X have a Binomial distribution? **No**
 - Why or why not? **Not a fixed number of trials**
 - What is the probability that the first time I roll a ‘6’ is on the third roll?

$$\mathbb{P}(X = 2) = \left(\frac{5}{6}\right)^2 \frac{1}{6}$$

Definition: Let X be the number of failures before the first success occurs in a sequence of independent Bernoulli trials with the probability of success on each trial begin p . Then X has the **Geometric distribution with parameter p** and we write

$X \sim \text{Geometric}(p)$.

Like the Binomial distribution, the Geometric distribution is defined in terms of a sequence of Bernoulli trials.

- The Binomial distribution counts the number of successes out of a fixed number of Bernoulli(p) trials.
- The Geometric distribution counts the **number trials before the first success occurs in a sequence of Bernoulli(p) trials**.

Warning note: In many texts (and in early versions of this course) the Geometric distribution is defined as the number of trials up to and including the first success. The change in definition results in all the formulae which follow being different.

6.3.1 Geometric probability mass function

If $X \sim \text{Geometric}(p)$, the p.m.f. of X is

$$f_X(x) = \mathbb{P}(X = x) = (1 - p)^x p \text{ for } x = 0, 1, 2, 3 \dots$$

Note: $\mathbb{P}(X = x) = \underbrace{(1 - p)^x}_{\text{need } x \text{ failures}} \times \underbrace{p}_{\text{final trial must be a success}}$

6.3.2 Geometric cumulative distribution function

The c.d.f. of X is $F_X(x) = \mathbb{P}(X \leq x) = 1 - (1 - p)^{x+1}$.

It follows that $\mathbb{P}(X > x) = (1 - p)^{x+1}$.

Why :

It is actually easier to start by thinking about $\mathbb{P}(X > x)$. Intuitively this means we must have $x + 1$ failures (what happens after that doesn't matter). thus $\mathbb{P}(X > x) = (1 - p)^{x+1}$ and the c.d.f formula follows.

6.3.3 Mean and Variance of the Geometric distribution

If $X \sim \text{Geometric}(p)$ then:

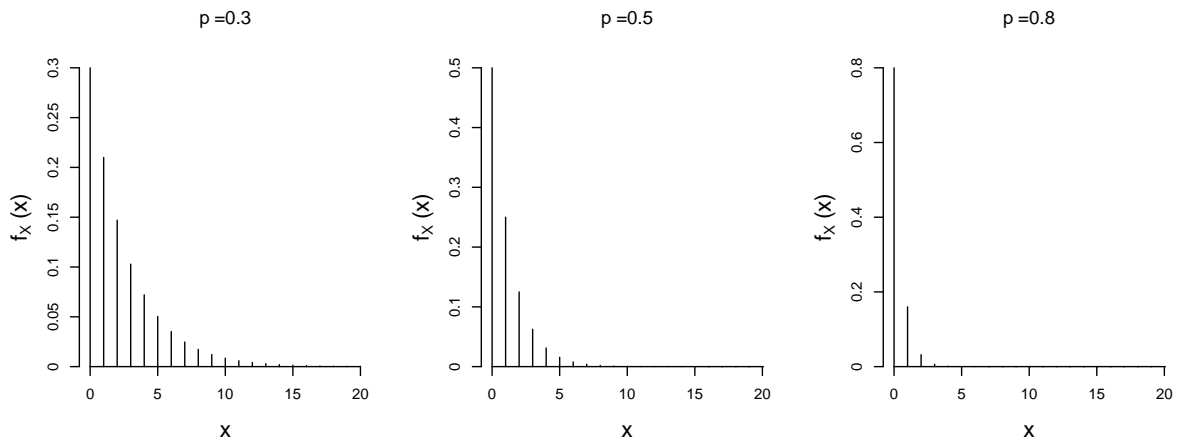
$$\mathbb{E}[X] = \frac{1 - p}{p} = \frac{q}{p}$$

$$\text{Var}(X) = \frac{1 - p}{p^2} = \frac{q}{p^2}, \text{ where } q = 1 - p$$

6.3.4 Shape of the Geometric distribution

The shape of the Geometric distribution depends upon the value of p . For small p , it is likely that there will be many failures before a success occurs, so the distribution has a long tail. For large p , a success is likely to occur almost immediately, so the distribution has a short tail. The geometric distribution is always positively skewed (right skewed).

Plots of the probability mass functions for $p = 0.3, 0.5, 0.9$ are shown below.



Example 6.8. In a version of the game Roulette, there are 37 numbered sectors $0, 1, 2, \dots, 36$ on a disk. The probability of success when betting on a single number is $1/37$. Assume that successive spins of the roulette wheel are independent of each other. Let X be the number of spins before your first win (each time betting on a single number).

What is an appropriate distribution for X

$$X \sim \text{Geometric}(p = \frac{1}{37})$$

What is the probability that you win on your first spin?

$$\mathbb{P}(X = 0) = (1 - p)^0 p = 1 \times \frac{1}{37} = \frac{1}{37}$$

How many times should you expect to lose before your first win?

$$\mathbb{E}[X] = \frac{1 - p}{p} = \frac{1 - \frac{1}{37}}{\frac{1}{37}} = 36$$

What is the probability you lose exactly 50 times before your first win?

$$\mathbb{P}(X = 50) = (1 - p)^x p = (1 - \frac{1}{37})^{50} \times \frac{1}{37} = 0.00687$$

If you lose on the first 3 spins, what is the probability that your first win occurs on your eighth spin?

$$\begin{aligned} \mathbb{P}(X = 7 | X \geq 3) &= \frac{\mathbb{P}(\{X = 7\} \cap \{X \geq 3\})}{\mathbb{P}(X \geq 3)} \\ &= \frac{\mathbb{P}(X = 7)}{\mathbb{P}(X > 2)} \\ &= \frac{(1 - \frac{1}{37})^7 \times \frac{1}{37}}{(1 - \frac{1}{37})^3} \\ &= (1 - \frac{1}{37})^4 \times \frac{1}{37} \\ &= 0.0242 \end{aligned}$$

Note: This is the same as $\mathbb{P}(X = (7 - 3)) = \mathbb{P}(X = 4)$ and illustrates the memoryless property of the geometric distribution.

See also Exercise 6.8.13.

6.4 Negative Binomial distribution

To win a tennis match a player must be the first to win two sets. Assume (probably incorrectly) that player A wins a set with probability 0.6, independent of the result of any previous sets. What is the probability that player A loses 1 set before winning the game?

How can this occur? **LWW or WLW**

So $\mathbb{P}(\text{Player A loses once before winning}) = 2 \times 0.4 \times 0.6^2 = 0.288$

In this case we had a sequence of Bernoulli trials but we could use neither the Binomial nor the Geometric distribution.

The Negative Binomial Distribution is used to model situations like this where we are counting the number failures before obtaining a fixed number of successes in a sequence of Bernoulli trials.

Definition: Let X be the number of failures before the k th success in a sequence of independent Bernoulli trials with the probability of success on each trial being p . Then X has the **negative Binomial distribution with parameters k and p** and we write $X \sim \text{NegBin}(k, p)$.

6.4.1 Negative Binomial probability mass function

If $X \sim \text{NegBin}(k, p)$, the p.m.f. of X is

$$f_X(x) = \mathbb{P}(X = x) = \binom{k+x-1}{x} p^k (1-p)^x \text{ for } x = 0, 1, 2, \dots$$

Explanation:

Each trial is independent with $\mathbb{P}(\text{success}) = p$

If the k th success is obtained after x failures then the k th success must occur on the $(k+x)$ th trial.

For this to occur we must first obtain $k-1$ successes and x failures, in any order, in the first $k+x-1$ trials. The probability of obtaining x failures (in any order) in the first $k+x-1$ trials can be calculated using the Binomial distribution. So

$$\mathbb{P}(x \text{ failures (in any order) in the first } k+x-1 \text{ trials}) = \binom{k+x-1}{x} p^{k-1} (1-p)^x$$

Note: In some formulations of the Negative Binomial distribution X is the number of trials to obtain r successes. As with the Geometric distribution this changes all the formulae.

6.4.2 Mean and Variance of the Negative Binomial distribution

If $X \sim \text{NegBin}(k, p)$ then:

$$\begin{aligned}\mathbb{E}[X] &= \frac{k(1-p)}{p} = \frac{kq}{p} \\ \text{Var}(X) &= \frac{k(1-p)}{p^2} = \frac{kq}{p^2}\end{aligned}$$

Proof: If $X \sim \text{NegBin}(k, p)$ then we can think of X as the sum of k Geometric random variables.

That is $X = Y_1 + Y_2 + \dots + Y_k$ where each $Y_i \sim \text{Geometric}(p)$.

Recall that this means that each Y_i is the number failures before the first (next) success this mean that $Y_1 + Y_2 + \dots + Y_k$ will be the same as **the total number of failures**.

From Section 6.3.3 we know that if $Y_i \sim \text{Geometric}(p)$ then $\mathbb{E}[Y_i] = \frac{q}{p}$.

Thus

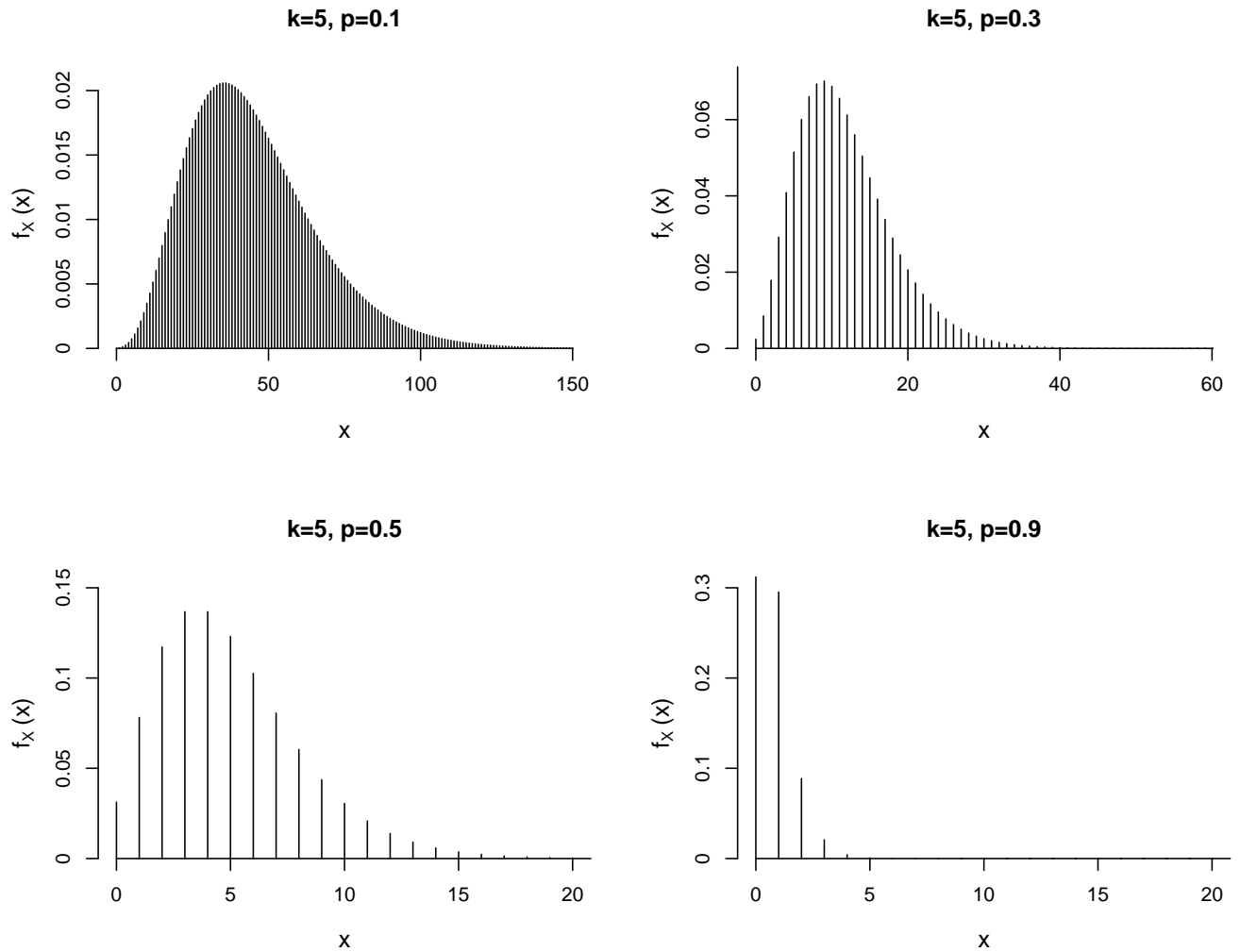
$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}[Y_1 + Y_2 + \dots + Y_k] \\ &= \mathbb{E}[Y_1] + \mathbb{E}[Y_2] + \dots + \mathbb{E}[Y_k] && \text{from Theorem 5.3} \\ &= \frac{q}{p} + \frac{q}{p} + \dots + \frac{q}{p} \\ &= \frac{kq}{p}\end{aligned}$$

The proof of the variance formula is similar, but, since it uses a formula from Chapter 8, it is left for you to try as an exercise after doing Example 8.5.

6.4.3 Shape of the Negative Binomial distribution

The shape of the negative Binomial distribution depends on the values of r and p .

The probability mass functions for $k = 5$ and various p are shown below.



Example 6.9. Suppose the probability of a successful missile launch is 0.9. Test launches are conducted until three successful launches are achieved.

1. If X is the number of failed launches conducted before the third successful launch is achieved, state an appropriate distribution for X .

$$X \sim \text{NB}(k = 3, p = 0.9)$$

2. What is the probability that exactly four failures will occur before the requirement of three successful launches is met?

$$\mathbb{P}(X = 4) = f_X(4) = \binom{3 + 4 - 1}{4} 0.9^3 (1 - 0.9)^4 = 0.00109$$

3. What are the mean and variance of the number of failed launches?

$$\mathbb{E}[X] = \frac{3 \times (1 - 0.9)}{0.9} = \frac{1}{3}$$

$$\text{Var}(X) = \frac{3 \times (1 - 0.9)}{0.9^2} = 0.3704$$

4. What is the probability that exactly five launches will be required?

Requiring five launches means $5 - 3 = 2$ failures so:

$$\mathbb{P}(X = 2) = \binom{3 + 2 - 1}{2} 0.9^3 (1 - 0.9)^2 = 0.04374$$

5. What is the probability that more than two launches will be required?

1 Why?!

6.5 Hypergeometric distribution

The hypergeometric distribution is another counting distribution.

Consider drawing marbles from a bag containing 70 red and 30 blue marbles, and recording whether or not the marble drawn is red. Let X be the number of red marbles drawn in 20 draws.

If we replace each marble in the bag after recording the colour then we have a sequence of Bernoulli trials and $X \sim \text{Binomial}(20, 0.7)$

Suppose however we do not return the marbles to the bag, i.e. we sample without replacement, then the probability of success will change after each draw. It then becomes easier to think of this probability in terms of the number of ways to draw x red marbles in 20 draws and the number of ways the 20 draws can be made.

There are $\binom{100}{20}$ ways to draw 20 marbles from 100.

Each draw is either a red marble or a blue marble. There are 70 possible red marbles and 30 possible blue marbles.

There are $\binom{70}{x}$ ways to draw x red marbles from the 70 red marbles.

The remaining $20 - x$ draws must be blue marbles .

There are $\binom{30}{20 - x}$ ways to draw $20 - x$ blue marbles from the 30 blue marbles.

The hypergeometric distribution is used to model situations like this where we are counting the number of “success” in n trials, sampling without replacement.

Definition: Let X be the number “successes” in n draws from a finite number of items N , of which M are “successes”. Then X has the **hypergeometric distribution with parameters n , M and N** and we write $X \sim \text{HYP}(n, M, N)$ or $X \sim \text{Hypergeometric}(n, M, N)$.

Note that: $n = 1, 2, 3, \dots, N$ and $M = 0, 1, 2, \dots, N$

If $X \sim \text{HYP}(n, M, N)$, the p.m.f. of X is

$$f_X(x) = \mathbb{P}(X = x) = \frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}} \text{ for } x = \max(0, n - N + M) \dots, \min(n, M)$$

Note:

There are $\binom{N}{n}$ ways to draw n items from N .

Each draw is either a “success” or a “failure”.

There are M possible “successes” and therefore $N - M$ possible “failures”.

There are $\binom{M}{x}$ ways to draw x “successes” from the M “successes”.

The remaining $n - x$ draws must be “failures”.

There are $\binom{N-M}{n-x}$ ways to draw $n - x$ “failures” from the $N - M$ “failures”.

We also need to think about what values it is possible for x to take on.

In most cases the minimum number of “successes” possible will be 0. However if n is greater than the number of possible “failures” then there must be at least $n - (N - M)$ “successes”.

Similarly in most cases the maximum number of “successes” possible will be n . However if n is greater than the number of possible “successes” then there can be at most M “successes”.

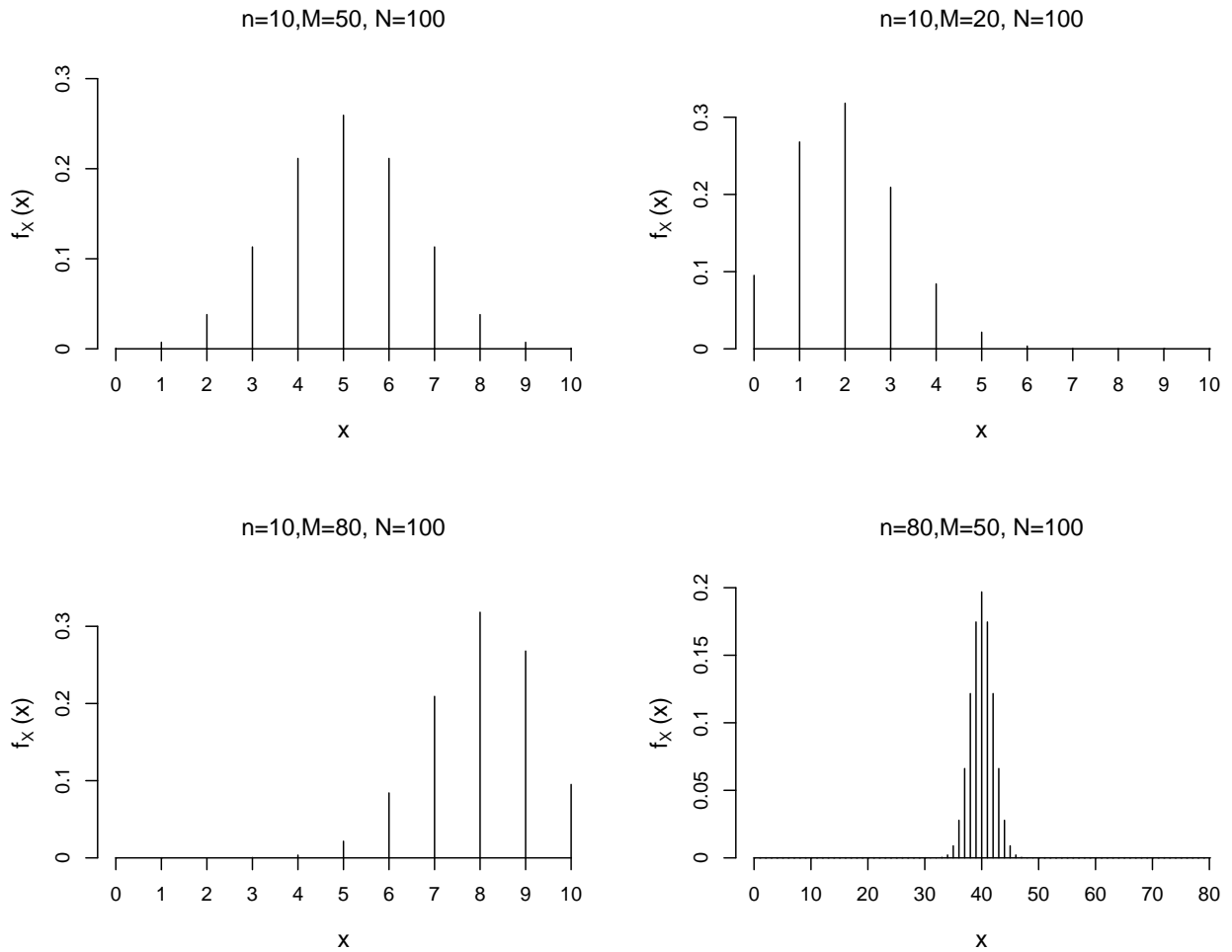
The mean and variance of the hypergeometric distribution can be shown to be:

$$\mathbb{E}[X] = \frac{nM}{N}$$

$$\text{Var}(X) = n \frac{M}{N} \left(1 - \frac{M}{N} \right) \frac{N-n}{N-1}$$

The shape of the hypergeometric distribution depends on the values of n , M and N .

The probability mass functions for $N = 100$ and various n and M are shown below.



Example 6.10. An application of the hypergeometric distribution is in deciding whether or not to accept a lot of manufactured items.

Suppose a shipment of 50 mechanical devices contains 42 good ones and 8 defective. An inspector selects five devices at random without replacement.

1. If X is the number of good devices selected, state an appropriate distribution to use to model X ?

$$X \sim \text{HYP}(n = 5, M = 42, N = 50)$$

2. What is the probability that exactly three good ones are selected?

$$\mathbb{P}(X = 3) = \frac{\binom{42}{3} \binom{50-42}{5-3}}{\binom{50}{5}} = 0.1517$$

3. What is the mean number of good ones selected?

$$\mathbb{E}[X] = \frac{5 \times 42}{50} = 4.2$$

4. What is the probability that at least one of the devices selected will be defective?

At least 1 defective means 4 or fewer are good so we want:

$$\mathbb{P}(X \leq 4) = 1 - \mathbb{P}(X = 5) = 1 - \frac{\binom{42}{5} \binom{50-42}{5-5}}{\binom{50}{5}} = 0.5985$$

6.5.1 Distinguishing the Hypergeometric distribution from the Binomial

Both the distributions count the number of successes in a fixed number of trials. So how do we decide which distribution to use?

A Binomial distribution assumes that we have independent trials with a constant probability of success. In practice this means that we are either sampling with replacement or we are sampling from an infinite (or at least very large) population. If we are sampling without replacement from a (small) finite population we should use the Hypergeometric distribution.

Example 6.11. Recall Example 6.5 where:

X = number of students with student loans $> \$10,000$.

n = number of trials = 6

and p = probability of success = 0.2.

We assumed that we were sampling from a large population of students so
 $X \sim \text{Binomial}(n = 6, p = 0.2)$

with p.m.f. $f_X(x) = \mathbb{P}(X = x) = \binom{6}{x} 0.2^x (1 - 0.2)^{6-x}$ for $x = 0, 1, \dots, 6$.

If instead we are choosing the 6 students from a class of 50 students, 20% of whom have a student loan $> \$10,000$. Then if Y = number of students with student loans $> \$10,000$.

$Y \sim \text{Hypergeometric}(N = 50, M = 0.2 \times 50 = 10, n = 6)$

1. What is the probability of getting (exactly) five students with student loans greater than \$10,000 out of a total of six students? Compare your answer to that of Example 6.5, where $\mathbb{P}(X = 5) = 0.001536$, what do you notice? Will this always be true?

$$\mathbb{P}(Y = 5) = \frac{\binom{10}{5} \binom{40}{1}}{\binom{50}{6}} = 0.00063$$

$\mathbb{P}(Y = 5)$ is smaller than $\mathbb{P}(X = 5)$, but this cannot always be the case.

2. Find $\mathbb{E}[X]$ and $\mathbb{E}[Y]$. What do you notice? Will this always be true?

$$\mathbb{E}[X] = np = 6 \times 0.2 = 1.2 \text{ and } \mathbb{E}[Y] = \frac{nM}{N} = \frac{6 \times 10}{50} = 1.2.$$

They are the same. This will always be true.

3. Find $\text{Var}(X)$ and $\text{Var}(Y)$. What do you notice? Will this always be true?

$$\text{Var}(X) = np(1 - p) = 6 \times 0.2 \times 0.8 = 0.96 \text{ and}$$

$$\text{Var}(Y) = n \frac{M}{N} \left(1 - \frac{M}{N}\right) \frac{N-n}{N-1} = 6 \times \frac{10}{50} \times \left(1 - \frac{10}{50}\right) \frac{50-6}{50-1} = 0.8620.$$

The variance for the situation using the hypergeometric is smaller. This will always be true.

6.6 Discrete Uniform distribution

Suppose I roll a fair 6-sided die. Let X be the number obtained on a single roll of the die then $\mathbb{P}(X = 1) = \mathbb{P}(X = 2) = \mathbb{P}(X = 3) = \mathbb{P}(X = 4) = \mathbb{P}(X = 5) = \mathbb{P}(X = 6) = \frac{1}{6}$.

The discrete Uniform distribution is used to model situations like this where it is reasonable to assume that all N values that a random variable can take have the same probability $\frac{1}{N}$.

If X has a discrete Uniform distribution with parameter N , the the p.m.f. of X is:

$$f_X(x) = \mathbb{P}(X = x) = \frac{1}{N} \text{ for } x = 1, 2, 3, \dots, N$$

We write $X \sim DU(N)$.

The mean and variance of the discrete Uniform distribution can be shown to be:

$$\begin{aligned}\mathbb{E}[X] &= \frac{N+1}{2} \\ \text{Var}(X) &= \frac{N^2-1}{12}\end{aligned}$$

Example 6.12. Suppose you have four cards in your hand one of each of the four suits (clubs, spades, hearts and diamonds). Let X be the number of random picks without replacement to get a heart.

1. Calculate $\mathbb{P}(X = 1)$
2. Calculate $\mathbb{P}(X = 2)$
3. Calculate $\mathbb{P}(X = 3)$
4. Calculate $\mathbb{P}(X = 4)$
5. What do you notice? What distribution could be used to model X

6.7 Poisson distribution

Example 6.13 (Customers at a fast food outlet). Suppose that customers arrive at an average rate of 2 per minute, independently of each other.

If X = number of customers to arrive in a 1-minute period, we can use the Poisson distribution to model X .

Definition: If X has a Poisson distribution with parameter λ , the p.m.f. of X is

$$f_X(x) = \mathbb{P}(X = x) = \frac{\lambda^x}{x!} e^{-\lambda} \text{ for } x = 0, 1, 2, \dots$$

We write $X \sim \text{Poisson}(\lambda)$.

The parameter λ is called the **intensity** of the Poisson distribution.

In the fast food example above:

- If X is the number of customers arriving in a 1-minute period then $\lambda = 2$. (2 per minute)
- If X is the number of customers arriving in a 5-minute period then $\lambda = 10$. (5×2 per 5 minute period)
- If X is the number of customers arriving in a 1-hour period then $\lambda = 120$. (60×2 per 60 minute period)

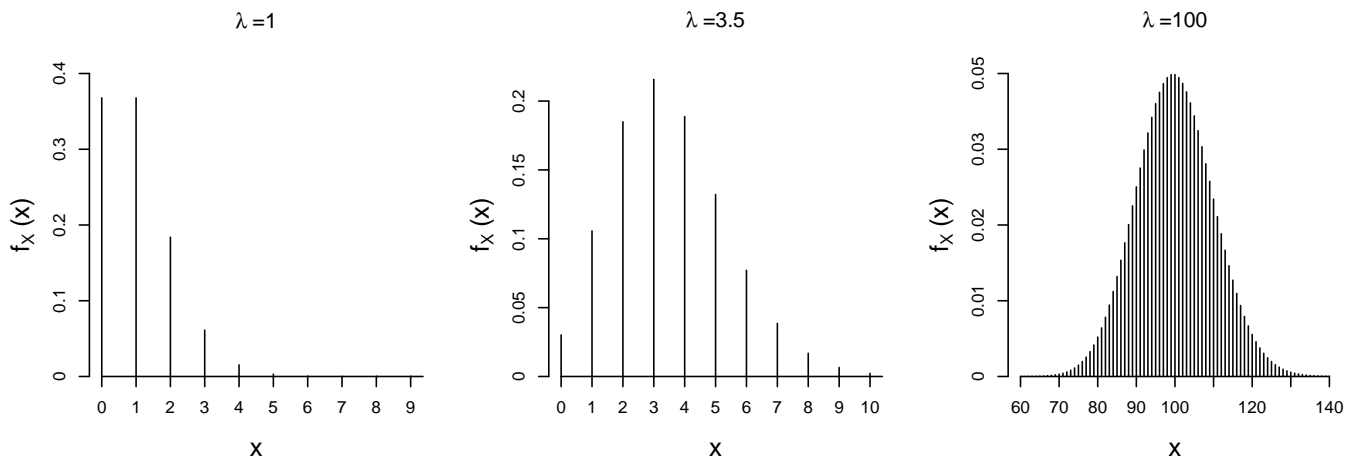
The mean and variance of the $\text{Poisson}(\lambda)$ distribution are both λ .

$$\mathbb{E}[X] = \text{Var}(X) = \lambda \text{ when } X \sim \text{Poisson}(\lambda)$$

Notes:

1. It makes sense for $\mathbb{E}[X] = \lambda$. If arrivals occur at a constant average rate of λ per unit time, then the mean of the number of arrivals to occur in one unit of time should indeed be λ .
2. The variance of the Poisson distribution increases with the mean (in fact, variance = mean). This is very often the case in real life: there is more uncertainty associated with larger numbers than with smaller numbers.
3. The shape of the Poisson distribution depends upon the value of λ . For small λ , the distribution has positive (right) skew. As λ increases, the distribution becomes more and more symmetrical, until for large λ it has the familiar bell-shaped appearance.

The probability mass functions for various λ are shown below.



4. Like the Binomial and Geometric distributions, the Poisson distribution is a distribution that arises in nature, through the so-called **Poisson process**. Roughly speaking, the Poisson process counts the **number of events occurring in a fixed time or space, when events occur independently and at a constant average rate**.

Example 6.14. Suppose that a website receives an average of 3 hits per minute. Assume that the number of hits per minute follows a Poisson distribution with $\lambda = 3$.

1. What is the probability of the website receiving exactly 3 hits in any given minute?

$$P(X = 3) = \frac{\lambda^x}{x!} e^{-\lambda} = \frac{3^3}{3!} e^{-3} = 0.224$$

2. What is the probability that the website receives no hits in one minute?

$$P(X = 0) = \frac{\lambda^x}{x!} e^{-\lambda} = \frac{3^0}{0!} e^{-3} = 0.0498$$

3. What is the probability that the website receives at least one hit every minute for 10 minutes?

$$P(X \geq 1)^{10} = (1 - P(X = 0))^{10} = (1 - 0.0498)^{10} = 0.9502^{10} = 0.600$$

4. Suppose that the website owner wants the probability of no hits in a minute to be 0.1 or less. If we assume that the number of hits per minute can be modelled using a Poisson distribution what would the average number of hits per minute need to be?

If $X \sim \text{Poisson}(\lambda)$ we want $\mathbb{P}(X = 0) \leq 0.1$. Thus

$$\begin{aligned}\mathbb{P}(X = 0) &= \frac{e^{-\lambda} \lambda^0}{0!} \leq 0.1 \\ \Rightarrow e^{-\lambda} &\leq 0.1 \\ \Rightarrow -\lambda &\leq \ln(0.1) \\ \Rightarrow \lambda &\geq 2.3\end{aligned}$$

Thus the average number of hits per minute would need to be at least 3.

Note: The Poisson process is a mathematically exact situation that will always result in a Poisson distribution. However, the Poisson distribution is also widely used as a ‘subjective model’ in situations that are not mathematically exact. Statisticians use subjective models when they need to describe the randomness in a situation that has no known mathematical formulation. Essentially, they are suggesting that the shape and variability of the distribution they are interested in is well captured by a Poisson distribution.

The difference between an exact model and a subjective model is important. Exact models, such as the Binomial distribution from Bernoulli trials, or the Poisson distribution from the Poisson process, are quite rare in real life; it is far more common for a subjective model to be required.

Example 6.15. Let X be the number of children of a randomly selected NZ woman. A reasonable subjective model for X might be $X \sim \text{Poisson}(\lambda = 2.5)$.

Despite this, the variance of the Poisson distribution is often *too small* to describe real-life situations adequately. In real life, the variance of a phenomenon often increases faster than the mean.

6.8 Exercises

6.8.1 Teams in a soccer competition are very similar in ability. The number of goals scored by a team in any match has a Poisson distribution with parameter 1, independent of the goals scored by the other team and independent of all other matches. In the whole competition, a total of 90 matches are played.

1. Find the probability that a fixed team scores no goals in a given match.
2. Given that a fixed team scores no goals in a given match, find the probability that they lose that match.
3. Given that a fixed team scores exactly 2 goals in a given match, find the probability that they win that match.
4. Give a formula for the probability that a given match ends in a draw.
5. What is the distribution of the number of matches in the competition in which no goals are scored?
6. What is the distribution of the number of matches that end in a draw?

6.8.2 Suppose you get a holiday job washing cars to help pay for your university education. The number of cars you wash in a day follows a Poisson distribution with $\lambda = 12$. You get paid \$25 when you arrive each day, and then \$6 for every car that you wash.

- (a) How much do you make each day, on average?
- (b) What is the probability that you make either \$49 or \$55 in one day?
- (c) What is the probability that you make more than \$37 in one day?
- (d) Suppose that the owner of the car wash gives you the option of receiving nothing when you arrive each day, but getting paid \$10 for each car that you wash. Assuming that your car washing distribution does not change, which option should you choose? Be sure to justify your answer (assume that you are trying to maximize the amount of money you make).

6.8.3 Suppose that births in a family are female with probability .48 and male with probability .52, independently of all previous births. A couple who are currently childless decide to continue having children until they have a daughter. Let X be the number of sons that they will have.

1. What is the distribution of X ?
2. What is the probability that the couple will have at least 5 children?

3. Find $\mathbb{E}[X]$.
4. A different childless couple (with the same birth probabilities) decide to keep having children until they have both a son *and* a daughter. Let Y be the number of children that they will have. Give an expression for $\mathbb{P}(Y = n)$ (hint: it is very helpful here to first condition on the gender of the first child!)

6.8.4 Suppose that a quality control inspector is responsible for checking for defective products at a computer chip manufacturing plant. A sample of 12 computer chips is taken from each batch of 1000, and tested for defects. The probability of a chip exhibiting a defect is 0.07.

- (a) What is the probability that no defects are observed in the sample of 12 chips?
- (b) What is the probability that less than three chips are found to be defective in the sample of 12?
- (c) If more than three chips in the sample of 12 are found to be defective, the entire batch of 1000 is rejected.
 - (i) What is the probability of rejecting a batch?
 - (ii) If five batches are produced each day, what is the probability that none are rejected?
 - (iii) What is the probability that no batches are rejected three days in a row?

6.8.5 At a small bar in West Lafayette, Indiana, there are three light beers (Busch, Coors and Budweiser) and two dark beers (Guinness and Bass) on tap. After a hard day at work two PhD students enter the bar and each randomly select a beer to drink.

- (a) What is the probability that the two students both select a light beer?
- (b) What is the probability that they both select the same beer?
- (c) What is the probability that they don't both select a dark beer?
- (d) Suppose that these same students return to the same bar each day for a week (5 days), and each randomly select one beer to drink.
 - (i) What is the probability that their beers will not match (i.e., on any given day, the two students do not each choose the same beer) on *any* day of the week?
 - (ii) What is the probability that their beers will match on at least two days?
 - (iii) What is the probability that the *color* of their beer (dark or light) will match *every* day of the week?

6.8.6 Suppose that a chocolate loving lecturer has \$12 to spend on chocolate. Her local store sells unlabelled blocks of chocolate for \$3 each, and she is trying to find chocolate with a certain flavour. Assume that the probability of an individual chocolate bar satisfying her taste is 0.2. Once she starts buying chocolate, she will continue until she finds the right flavour, or until she runs out of money.

- (i) What is the probability that she finds the right flavour on her third purchase?
- (ii) What is the probability that she runs out of money before finding the right flavour?
- (iii) How much should she expect to pay to find the correct flavour?
- (iv) How much money would she need to bring to have a 90% chance of finding the correct flavour?

6.8.7 An electronics store has decided to import digital cameras from a new supplier. Each day it receives five cameras from the supplier, and each camera is tested before being displayed in the retail store. The probability that a camera fails the test is equal to 0.17, and is independent of whether or not the other cameras fail.

- (a) What is the probability that exactly two of the five cameras fail on a particular day?
- (b) On average, how many of the five cameras do we expect to fail?
- (c) If we see more than one failure we reject that day's shipment. What is the probability of rejecting the shipment?
- (d) Assuming independence between days, what is the probability that shipments are rejected on both the second and the third days?
- (e) How many days do we expect to wait before rejecting a shipment?

6.8.8 A checkout operator at a busy service station serves an average of one customer every minute. Let X be the number of customers served in a particular minute.

- (a) What is an appropriate distribution for X ?
- (b) What is the probability that no customers are served in any given minute?
- (c) What is the probability that two or more customers are served in any given minute?
- (d) Suppose that a new checkout operator starts his shift. What is the probability that two minutes will pass with no customers and then at least one customer is served in the third minute?

- (e) Intuitively, how long do you think the new checkout operator has to wait (on average) at the start of a shift before serving his first customer? (We will be able to answer this question more satisfactorily later in the course.)

6.8.9 A simple gambling game involves dealing two cards (without replacement) from a deck of 52 cards, of which half are red, and half are black. If you get two red cards, or two black cards, you win \$5. If you get a mix of colours (i.e., one black card and one red card), you lose \$5.

- (a) What is the probability of getting two red cards?
- (b) What is the probability of winning \$5?
- (c) What is the expected value of this game?
- (d) What is the probability that you lose this game *at least twice* before you win for the first time?
- (e) What is the probability that you lose exactly four times out of your first ten games?

6.8.10 A police officer operating a speed camera sees an average of three speeding cars go past every minute. Each speeding car is photographed by the camera. Let X be the number of cars photographed in a 60 second period.

- (a) What is an appropriate distribution for X ?
- (b) What is the probability that 0 cars are photographed in any given minute?
- (c) What is the probability that more than three cars are photographed in the space of one minute?
- (d) Suppose that out of all the drivers caught speeding, 70% of them are male, and that a random sample of 10 speeding drivers is taken (you may assume that the population size is infinite).
 - (i) What is the probability that 8 members of the sample are male?
 - (ii) What is the probability that 0 members of the sample are female?
 - (iii) How many females would we expect to get in the sample of 10?

6.8.11 The popular Australian gambling game “Two-up” (which can only be legally played on Anzac Day), involves repeatedly tossing two coins. A player tosses two coins until one of three outcomes occurs: two heads, two tails, or five “mixed” (one head and one tail) tosses in a row. If two heads are thrown, the player wins \$5, and their turn is over. If two tails are thrown, or five mixed tosses occur, then the player loses \$5 and their turn is over.

- (a) Draw a probability tree to represent this game.

- (b) What is the probability of winning?
- (c) What is the expected value of the game?
- (d) What would the game have to pay for a win (i.e., two heads) to make it a fair game? (Recall that a fair game has an expected value of zero).
- (e) What is the probability that a player losses three times before their first win?
- (f) How many games should a player expect to lose before their first win?

6.8.12 25 scientists in different countries independently collect data with the aim of establishing a relationship between banana consumption and elbow cancer in humans. They each use a statistical test that has the property that: in 5% of cases where there is truly no relationship between the quantities of interest, the test will suggest that there is one. Suppose that there is truly no relationship between banana consumption and elbow cancer in humans.

- (a) What is the distribution of the number of scientists (out of 25) whose test suggests that there is a relationship?
- (b) Find the probability that at least one scientist finds a relationship.
- (c) Find the probability that at least two scientists find a relationship.
- (d) Suppose that a scientist is convinced that there is a relationship and so will repeat the study (with independent data collection each time) once each year until the test suggests a relationship. What is the expected number of studies that the scientist performs?

6.8.13 During the lunch hour, customers enter a bank as a Poisson process. On average in a lunch hour 10 customers enter the bank.

- (a) In this course we study a number of “named” discrete distributions (e.g. Binomial, Geometric, Poisson) with certain parameters. In each case below, match the appropriate distribution with the random variable described, and give the appropriate parameter(s).
 - (i) The number of arrivals, N , in one given lunch hour.
 - (ii) The number of lunch hours in a business week (Monday to Friday) in which at least one customer arrives.
 - (iii) The number of business days before at least one customer arrives during the lunch hour.
- (b) Based on your answers above, find the probability that no customers arrive in the lunch hour for an entire business week.
- (c) One of the distributions (Binomial, Geometric or Poisson) has a “memoryless” property of the form

$$\mathbb{P}(X > n + m | X > n) = \mathbb{P}(X > m - 1).$$

Which one is it? Prove this (for general value(s) of the parameter(s) of that distribution).

6.8.14 For all questions below, assume that we are working with a binomial distribution with parameters $n = 8$ and $p = 0.2$. Immediately before generating the random sample in part (c), please issue the following command:
`set.seed(123)`

- (a) What is the probability of observing exactly two successes for a random variable X following the binomial distribution described above?
- (b) What is the probability of observing less than two successes?
- (c) Use the `rbinom` command to generate 100 samples from a binomial distribution with $n = 8$ and $p = 0.2$. List the first five samples.
- (d) Use the `hist` command to produce a histogram of the random sample.
- (e) What proportion of the samples were equal to two? How does this compare to the probability of getting two successes (calculated above in part (a))?
- (f) What proportion of the samples were less than two? How does this compare to the probability of less than two successes (calculated above in part (b))?
- (g) Why would you expect your answers to parts (e) and (f) to be (at least slightly) different to your answers to parts (a) and (b)?
- (h) Calculate the mean and variance of your sample.

7 Joint and conditional distributions

By the end of this chapter you should be able to:

- compute joint probability mass functions
- recognize pairs of random variables as independent or dependent
- compute conditional and marginal distributions

7.1 Joint distributions

Until now, we have considered random variables one at a time: the expected value and variance of a random variable X , the probability that X takes the value 0, and so on. These properties depend only on the distribution of X (and not on the details of the sample space, for instance) and can therefore be computed based on knowing the p.m.f. $f_X(x)$ or the c.d.f. $F_X(x)$.

We often wish to talk about the *joint distribution* of two or more random variables obtained from the outcome of a single random experiment. For instance, choose a person at random and measure their height and their weight. These random variables are related: a taller person is more likely to be heavier, but exactly how much heavier is unpredictable. To obtain a full description of this pair of variables, it is not enough to know the distribution of each variable separately. In this chapter, we will see how to describe the relationship between two random variables. We will mostly be looking at examples with two random variables, but the ideas can easily be extended to more than two.

Definition: The **joint probability mass function** for a pair of discrete random variables X, Y is given by

$$f_{X,Y}(x, y) = \mathbb{P}(X = x \text{ and } Y = y) = \mathbb{P}(X = x, Y = y) = \mathbb{P}(\{X = x\} \cap \{Y = y\})$$

for all possible values x and y of X and Y .

Note: in the joint p.m.f. $f_{X,Y}(x, y)$, the value $f_{X,Y}(x, y)$ is a probability, i.e., a number between 0 and 1. The arguments, x and y , are constant numbers that we can choose arbitrarily, depending on what we want to know. Usually x is chosen to be one of the possible values for X , and y is chosen to be one of the possible values for Y . (However, (x, y) might not be a possible value for (X, Y) , so $f_{X,Y}(x, y)$ could be zero or non-zero.) The values of the random variables X and Y do not influence $f_{X,Y}(x, y)$ – $f_{X,Y}(x, y)$ is a probability, not a random variable – but the joint distribution of X and Y determines what the joint p.m.f. is as a function of x and y .

Theorem 7.1. Let X, Y be discrete random variables.

- (a) Let $g(x, y)$ be a real-valued function defined for all possible values of X and Y . Then the expected value of the random variable $g(X, Y)$ is

$$\mathbb{E}(g(X, Y)) = \sum_x \sum_y g(x, y) f_{X,Y}(x, y) = \sum_x \sum_y g(x, y) \mathbb{P}(X = x, Y = y)$$

For instance, with $g(x, y) = xy$,

$$\mathbb{E}(XY) = \sum_x \sum_y xy f_{X,Y}(x, y) = \sum_x \sum_y xy \mathbb{P}(X = x, Y = y)$$

- (b) Let A be an event defined only in terms of X and Y , i.e., an event of the form $A = \{\text{the values } X, Y \text{ satisfy condition } \mathcal{C}\}$. Then the probability of A is

$$\mathbb{P}(A) = \sum_{x,y: x,y \text{ satisfy condition } \mathcal{C}} f_{X,Y}(x, y).$$

Theorem 7.1 says that any probability or expectation based on only X and Y can be computed in terms of the joint p.m.f. Just as in Section 5.5, the method can be summarised as: “The probability of obtaining particular values, times the corresponding value of the thing inside the expectation, summed over all possible values.”

Since the joint p.m.f. completely describes the distribution of X and Y together, we can use it to calculate the p.m.f. for each random variable separately.

Definition: The **marginal** probability mass functions for X and Y are given by

$$\begin{aligned} f_X(x) &= \mathbb{P}(X = x) = \sum_y f_{X,Y}(x, y) && \text{for all values } x, \\ f_Y(y) &= \mathbb{P}(Y = y) = \sum_x f_{X,Y}(x, y) && \text{for all values } y. \end{aligned}$$

The marginal p.m.f.’s are the same as the (ordinary) p.m.f.’s defined in Section 5.2. The word marginal is only to emphasise that we want to consider one variable separately.

Note that knowing the marginal p.m.f.’s $f_X(x)$ and $f_Y(y)$ only is not enough to find out expectations involving both random variables, such as $\mathbb{E}(XY)$ (except when we have extra information about X and Y , see Section 7.2).

Example 7.1. Let X and Y be the random variables with joint p.m.f.

$$f_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y) = \begin{cases} 0.2 & \text{if } x = 1, y = 3, \\ 0.4 & \text{if } x = 1, y = 6, \\ 0.1 & \text{if } x = 2, y = 3, \\ 0 & \text{if } x = 2, y = 6, \\ 0.1 & \text{if } x = 3, y = 3, \\ 0.2 & \text{if } x = 3, y = 6, \\ 0 & \text{otherwise.} \end{cases}$$

Find the marginal p.m.f.'s $f_X(x)$ and $f_Y(y)$, and compute $\mathbb{P}(X + Y \text{ is even})$ and $\mathbb{E}(XY)$.

In this example, there is no apparent pattern in the values of the p.m.f. $f_{X,Y}(x, y)$ as a function of (x, y) . It is more convenient to express the information in $f_{X,Y}(x, y)$ in a table:

$f_{X,Y}(x, y)$		y		
		3	6	
x	1	0.2	0.4	0.6
	2	0.1	0	0.1
	3	0.1	0.2	0.3
		0.4	0.6	1

Using Theorem 7.1,

$$\begin{aligned} \mathbb{P}(X + Y \text{ is even}) &= \sum_{x,y: x+y \text{ is even}} f_{X,Y}(x, y) = f_{X,Y}(1, 3) + f_{X,Y}(2, 6) + f_{X,Y}(3, 3) \\ &= 0.2 + 0 + 0.1 = 0.3 \end{aligned}$$

and

$$\mathbb{E}(XY) = (1)(3)(0.2) + (1)(6)(0.4) + (2)(3)(0.1) + (2)(6)(0) + (3)(3)(0.1) + (3)(6)(0.2) = 8.1.$$

The marginal p.m.f.'s can be written in the margins of the table for $f_{X,Y}(x, y)$, above, or written out explicitly as

$$f_X(x) = \begin{cases} 0.6 & \text{if } x = 1, \\ 0.1 & \text{if } x = 2, \\ 0.3 & \text{if } x = 3, \\ 0 & \text{otherwise,} \end{cases} \quad f_Y(y) = \begin{cases} 0.4 & \text{if } y = 3, \\ 0.6 & \text{if } y = 6, \\ 0 & \text{otherwise.} \end{cases}$$

The joint p.m.f. for a pair of random variables works almost exactly like the p.m.f.'s for one random variable as defined in Section 5.2. That is, a joint p.m.f. value such as $f_{X,Y}(3, 5)$ is the answer to a question – how probable is it that X will equal 3 and Y will equal 5, simultaneously? – and the joint p.m.f. $f_{X,Y}(x, y)$ is the “answer key” that specifies the answers to all possible questions where 3 is replaced by x and 5 is replaced by y . We can think of think of $f_{X,Y}(x, y)$ as the p.m.f. for a random pair (X, Y) , where (x, y) represents a possible value (or an impossible value, for that matter) of the random pair (X, Y) .

Example 7.2. Roll a fair die twice, and let X and Y be the numbers showing on the first and second rolls. Let

$$f_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y) = \begin{cases} \frac{1}{36} & \text{if } x, y \in \{1, 2, 3, 4, 5, 6\}, \\ 0 & \text{otherwise.} \end{cases}$$

be the joint probability mass function. Find the marginal p.m.f.'s $f_X(x)$ and $f_Y(y)$, and compute $\mathbb{P}(X + Y \geq 11)$ and $\mathbb{E}(XY)$.

[The formula for the joint p.m.f. $f_{X,Y}(x, y)$ reflects our idea that any valid pair of rolls should be equally likely, and invalid roll values are not allowed.]

To compute the marginal p.m.f. $f_X(x)$, note that if $x \notin \{1, 2, 3, 4, 5, 6\}$ then $f_{X,Y}(x, y) = 0$ for all values of y . If $x \in \{1, 2, 3, 4, 5, 6\}$ then $f_{X,Y}(x, y)$ has the value $1/36$ for exactly 6 values of y , and is 0 otherwise. So

$$f_X(x) = \mathbb{P}(X = x) = \begin{cases} \sum_{y=1}^6 \frac{1}{36} & \text{if } x \in \{1, 2, 3, 4, 5, 6\}, \\ 0 & \text{otherwise} \end{cases} = \begin{cases} \frac{6}{36} = \frac{1}{6} & \text{if } x \in \{1, 2, 3, 4, 5, 6\}, \\ 0 & \text{otherwise.} \end{cases}$$

The same argument gives $f_Y(y) = 1/6$ if $y \in \{1, 2, 3, 4, 5, 6\}$ and $f_Y(y) = 0$ otherwise. This result agrees with our intuition because, regardless of whether we make additional rolls, X and Y are individually the results of rolling one die.

We can also represent this calculation in tabular form:

$f_{X,Y}(x, y)$		y						
		1	2	3	4	5	6	
x	1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
	2	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
	3	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
	4	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
	5	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
	6	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
		$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1

The entries in the middle of the table are the values of the joint p.m.f.; the “0 otherwise” values are assumed. The right and bottom margins show the marginal distributions of X and Y , respectively. To verify our calculations, it is worth checking that we get a sum of 1 from the values in the middle and from the values in each margin.

To compute $\mathbb{P}(X + Y \geq 11)$, we apply Theorem 7.1. We only need to consider the values x, y for which $f_{X,Y}(x, y) \neq 0$, i.e., $x, y \in \{1, 2, 3, 4, 5, 6\}$. Of these, only $(x, y) = (5, 6), (6, 5)$, and $(6, 6)$ satisfy the condition $x + y \geq 11$, so

$$\mathbb{P}(X + Y \geq 11) = f_{X,Y}(5, 6) + f_{X,Y}(6, 5) + f_{X,Y}(6, 6) = \frac{1}{36} + \frac{1}{36} + \frac{1}{36} = \frac{3}{36} = \frac{1}{12}.$$

To compute $\mathbb{E}(XY)$ using Theorem 7.1, we again restrict to x, y for which $f_{X,Y}(x, y) \neq 0$. Then

$$\begin{aligned}\mathbb{E}(XY) &= \sum_{x,y} xy f_{X,Y}(x, y) = \sum_{x=1}^6 \sum_{y=1}^6 xy \cdot \frac{1}{36} \\ &= \frac{1}{36} (1 + 2 + 3 + 4 + 5 + 6 + 2 + 4 + 6 + \cdots + 24 + 30 + 36) \\ &= \cdots = \frac{441}{36} = \frac{49}{4}\end{aligned}$$

after a somewhat lengthy calculation. This calculation can be performed somewhat more efficiently by using the formula

$$\sum_{i=1}^n i = \frac{n(n+1)}{2}$$

twice, along with the rules for summation. In Section 7.2, we will see a much simpler method for finding this expectation.

Example 7.3. Choose a point uniformly at random from among the origin and the four closest integer-valued points, i.e., the points $(0, 0)$, $(1, 0)$, $(0, 1)$, $(-1, 0)$, $(0, -1)$. Let X, Y be the x - and y -coordinates of the point chosen.

- (a) Find the joint and marginal p.m.f.'s for X and Y .
- (b) Compute $\mathbb{P}(X = Y)$ and $\mathbb{E}(XY)$.

Since there are 5 equally likely possibilities

$$f_{X,Y}(x, y) = \begin{cases} \frac{1}{5} & \text{if } (x, y) = (0, 0), (1, 0), (0, 1), (-1, 0) \text{ or } (0, -1), \\ 0 & \text{otherwise.} \end{cases}$$

The marginal p.m.f.'s here are as follows:

$$\begin{aligned} f_X(-1) &= \mathbb{P}(X = -1) = \mathbb{P}(X = -1, Y = 0) = f_{X,Y}(-1, 0) = \frac{1}{5}, \\ f_X(0) &= \mathbb{P}(X = 0) = f_{X,Y}(0, 1) + f_{X,Y}(0, 0) + f_{X,Y}(0, -1) = \frac{3}{5}, \\ f_X(1) &= \mathbb{P}(X = 1) = f_{X,Y}(1, 0) = \frac{1}{5}. \end{aligned}$$

and $f_X(x) = 0$ otherwise. Similarly, we find that

$$f_Y(-1) = \frac{1}{5}, \quad f_Y(0) = \frac{3}{5}, \quad f_Y(1) = \frac{1}{5}.$$

In tabular form,

$f_{X,Y}(x, y)$		y			
		-1	0	1	
x	-1	0	$\frac{1}{5}$	0	$\frac{1}{5}$
	0	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{3}{5}$
	1	0	$\frac{1}{5}$	0	$\frac{1}{5}$
		$\frac{1}{5}$	$\frac{3}{5}$	$\frac{1}{5}$	1

Using Theorem 7.1,

$$\mathbb{P}(X = Y) = \sum_{x,y: x=y} f_{X,Y}(x, y) = f_{X,Y}(0, 0) = \frac{1}{5}$$

and

$$\mathbb{E}(XY) = \sum_{x,y} xy f_{X,Y}(x, y) = \frac{1}{5} \left((0)(0) + (1)(0) + (0)(1) + (-1)(0) + (0)(-1) \right) = 0.$$

In this example, the marginal p.m.f.'s f_X and f_Y are the same, so the distribution of X is the same as the distribution of Y . However, the random variables X and Y are not the same, i.e., $\mathbb{P}(X = Y) \neq 1$.

Example 7.4. Draw two cards from a standard deck. Set X to be 1 if the first card is an Ace, and set X to be 0 otherwise. Set Y to be 1 if the second card is an Ace, and set Y to be 0 otherwise. Find the joint and marginal p.m.f.'s for X and Y .

The possible values of X and Y are 0 and 1, so we will have $f_{X,Y}(x,y) = 0$ unless $x,y \in \{0,1\}$. Consider first $x = y = 1$. The event $\{X = 1, Y = 1\}$ is the event that both the cards drawn are Aces. Conditioning on the result of the first draw,

$$\begin{aligned}\mathbb{P}(X = 1, Y = 1) &= \mathbb{P}(\text{both cards are Aces}) \\ &= \mathbb{P}(\text{first card is an Ace})\mathbb{P}(\text{second card is an Ace} \mid \text{first card is an Ace}) \\ &= \mathbb{P}(\text{choose one of 4 Aces out of 52 cards}) \\ &\quad \cdot \mathbb{P}(\text{choose one of the 3 remaining Aces out of 51 remaining cards}) \\ &= \frac{4}{52} \cdot \frac{3}{51} = \frac{4 \cdot 3}{52 \cdot 51}.\end{aligned}$$

[No additional understanding is gained from multiplying out $4 \cdot 3 = 12$ and $52 \cdot 51 = 2652$; on the contrary, it makes the pattern rather less apparent.]

Similarly

$$\begin{aligned}\mathbb{P}(X = 0, Y = 0) &= \mathbb{P}(\text{choose one of 48 non-Aces out of 52 cards}) \\ &\quad \cdot \mathbb{P}(\text{choose one of the 47 remaining non-Aces out of 51 remaining cards}) \\ &= \frac{48}{52} \cdot \frac{47}{51} = \frac{48 \cdot 47}{52 \cdot 51}, \\ \mathbb{P}(X = 1, Y = 0) &= \mathbb{P}(\text{choose one of 4 Aces out of 52 cards}) \\ &\quad \cdot \mathbb{P}(\text{choose one of the 48 non-Aces out of 51 remaining cards}) \\ &= \frac{4}{52} \cdot \frac{48}{51} = \frac{4 \cdot 48}{52 \cdot 51}.\end{aligned}$$

By repeating the last argument, or by symmetry, $\mathbb{P}(X = 0, Y = 1) = \frac{48 \cdot 4}{52 \cdot 51}$ also. Adding to find the marginal distributions,

$$f_Y(0) = f_{X,Y}(0,0) + f_{X,Y}(1,0) = \frac{48 \cdot 47 + 4 \cdot 48}{52 \cdot 51} = \frac{48 \cdot 51}{52 \cdot 51} = \frac{48}{52},$$

and similarly $f_X(0) = 48/52$, $f_X(1) = f_Y(1) = 4/52$. In tabular form,

$f_{X,Y}(x,y)$		y		
		0	1	
x	0	$\frac{48 \cdot 47}{52 \cdot 51}$	$\frac{48 \cdot 4}{52 \cdot 51}$	$\frac{48}{52}$
	1	$\frac{4 \cdot 48}{52 \cdot 51}$	$\frac{4 \cdot 3}{52 \cdot 51}$	$\frac{4}{52}$
		$\frac{48}{52}$	$\frac{4}{52}$	1

Note that the marginal distributions for X and Y are the same. This must be so (assuming a properly shuffled deck) because the simplified experiments “take the top card” and “take the card one below the top card” are equivalent.

7.2 Independent random variables

Definition: Two discrete random variables X and Y are *independent* if and only if

$$(a) \quad \mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y) \quad \text{for all } x, y.$$

An equivalent statement is

$$(b) \quad f_{X,Y}(x, y) = f_X(x)f_Y(y) \quad \text{for all } x, y.$$

If either of these equivalent statements fail (for some pair of values x, y) then X and Y are *dependent*.

Several discrete random variables X_1, \dots, X_N are (*mutually*) *independent* if and only if $f_{X_1, \dots, X_N}(x_1, \dots, x_n) = f_{X_1}(x_1) \cdots f_{X_N}(x_N)$ for all values of x_1, \dots, x_n .

Example 7.5 (Example 7.1 continued). Let X and Y have the joint p.m.f. specified in Example 7.1. Then X and Y are **dependent** because, for instance,

$$f_{X,Y}(1, 3) = 0.2 \neq (0.6)(0.4) = f_X(1)f_Y(3)$$

Example 7.6 (Example 7.2 continued). Let X and Y (the dice rolls) be the random variables with joint p.m.f.

$$f_{X,Y}(x, y) = \begin{cases} \frac{1}{36} & \text{if } x, y \in \{1, 2, 3, 4, 5, 6\}, \\ 0 & \text{otherwise.} \end{cases}$$

Are X and Y independent random variables according to the definition?

We already computed $f_X(x) = f_Y(y) = \frac{1}{6}$ for all $x, y \in \{1, 2, 3, 4, 5, 6\}$, with $f_X(x) = f_Y(y) = 0$ otherwise.

If $x, y \in \{1, 2, 3, 4, 5, 6\}$ then $f_{X,Y}(x, y) = \frac{1}{36} = \frac{1}{6} \cdot \frac{1}{6} = f_X(x)f_Y(y)$.

Otherwise, one of x, y is not a possible value, so either $f_X(x)$ or $f_Y(y)$ must be 0. In this case $f_X(x)f_Y(y)$ and $f_{X,Y}(x, y)$ are both 0.

Thus $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ for all $x, y \in \{1, 2, 3, 4, 5, 6\}$, so X and Y are *independent*.

[Generally, when checking the definition of independence, it is only necessary to check $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ for possible values of x, y , i.e., for values of x, y such that $f_X(x) > 0$ and $f_Y(y) > 0$. If $f_X(x) = 0$, for instance, then automatically $f_{X,Y}(x, y) = 0$ and the equation $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ becomes the true equation $0 = 0$.]

Example 7.7 (Example 7.3 continued). Choose (X, Y) to be one of the 5 points $(0, 0), (1, 0), (0, 1), (-1, 0), (0, -1)$ with equal probability. Are X and Y independent or dependent?

We previously computed $f_X(0) = f_Y(0) = 3/5$, whereas $f_{X,Y}(0, 0) = 1/5$, and

$$f_{X,Y}(0, 0) = 1/5 \neq f_X(0)f_Y(0) = 9/25.$$

Even one pair of values x, y where $f_{X,Y}(x, y) \neq f_X(x)f_Y(y)$ is enough to show that X and Y are *dependent*.

In this case, a more striking counterexample is available. Both $x = 1$ and $y = -1$ are possible values for X and Y individually (i.e., $f_X(1)$ and $f_Y(-1)$ are positive) but the event $\{X = 1 \text{ and } Y = -1\}$ cannot occur (i.e., $f_{X,Y}(1, -1) = 0$). So $f_{X,Y}(x, y)$ cannot equal $f_X(x)f_Y(y)$ when $x = 1, y = -1$, and this is also enough to show that X and Y are *dependent*.

Example 7.8 (Example 7.4 continued). Draw two cards from a standard deck. Set X to be 1 if the first card is an Ace, and set X to be 0 otherwise. Set Y to be 1 if the second card is an Ace, and set Y to be 0 otherwise. Are X and Y independent or dependent?

We have

$$f_{X,Y}(1, 1) = \frac{4 \cdot 3}{52 \cdot 51} \neq f_X(1)f_Y(1) = \frac{4 \cdot 4}{52 \cdot 52}.$$

So X and Y are *dependent*.

Consequences of independence

Theorem 7.2. Let X and Y be independent random variables, and let g and h be real-valued functions defined for all possible values of X and Y , respectively. Then:

- (i) The random variables $g(X)$ and $h(Y)$ are independent.
- (ii) $\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y))$
- (iii) Let A be an event depending only on X , and B be an event depending only on Y . Then A and B are independent events.

Notes:

- If we specify the (marginal) p.m.f.'s $f_X(x), f_Y(y)$ of two random variables X, Y , and if we specify that X and Y are independent, then the *joint* distribution of X and Y is completely determined by the formula $f_{X,Y}(x, y) = f_X(x)f_Y(y)$.
- We can always create new random variables X_1, \dots, X_N such that X_1, \dots, X_N are independent and have any specified marginal distributions.¹
- Statement (i) generalises to say that if X_1, \dots, X_N are independent and $1 \leq n < N$, then $X = g(X_1, X_2, \dots, X_n)$ and $Y = h(X_{n+1}, X_{n+2}, \dots, X_N)$ are independent.
- If we specify the p.m.f.'s $f_{X_1}(x_1), f_{X_2}(x_2)$ but we do not specify that X_1 and X_2 are independent, then there are (in general) infinitely possible joint distributions for (X_1, X_2) .

Example 7.2/7.6 revisited: Previously, we set the joint p.m.f. for two dice rolls to be $f_{X,Y}(x, y) = 1/36$ for $x, y \in \{1, 2, 3, 4, 5, 6\}$, and we used this definition to find the marginal distributions and to check independence. A better way to express our views about dice rolls would be to specify that $f_X(x) = f_Y(y) = 1/6$ for $x, y \in \{1, 2, 3, 4, 5, 6\}$ and to specify that X and Y should be independent.

The equation $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ then determines the joint p.m.f. We can also use Theorem 7.2 to compute $\mathbb{E}(XY)$ more simply:

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y) = \left(\frac{6+1}{2}\right)\left(\frac{6+1}{2}\right) = \frac{49}{4}$$

since the marginal distribution of both X and Y is the discrete uniform distribution $\text{DU}(6)$.

¹The precise meaning is that, if f_{X_1}, \dots, f_{X_N} are any given p.m.f.'s, we may produce a sample space S' , usually larger and more complicated than the original sample space we might have had in mind, and random variables X'_1, \dots, X'_N defined on S' , such that X'_1, \dots, X'_N are mutually independent with marginal p.m.f.'s f_{X_1}, \dots, f_{X_N} . In short, we can always introduce new randomness that is independent of everything else, without changing the original probabilities. In practice, the fact that this required changing the sample space is usually ignored.

7.3 Conditional distributions

Consider a random experiment and a measurement, i.e., a random variable, X , depending on the result. Before the experiment is performed, our assessment of the possible results is encoded in the distribution of X – e.g., via the probabilities $(f_X(x), x \in \mathbb{R})$ in the p.m.f.

Suppose we are now given partial information about the experiment – we are told that the event A occurs. Based on this new information, we will revise our assessment of the possible results. Specifically, we will replace the probabilities we originally calculated by *conditional probabilities* given A . This means that our ideas about X must also change – we must assess the *conditional distribution* of X given A .

Recall also that $\mathbb{P}(X = x)$ is the probability of the event that the outcome s satisfies $X(s) = x$, i.e. the probability of the event $\{s \in S : X(s) = x\}$. Then we can handle things like $\mathbb{P}(X = x | A)$, which is the probability that the random variable X is equal to x , given that A occurs:

$$\mathbb{P}(X = x | A) = \frac{\mathbb{P}(\{s : X(s) = x\} \cap A)}{\mathbb{P}(A)}.$$

Example 7.9. Roll a fair die, and let X be the number showing on the die.

Let A be the event that the number is even.

What is the conditional probability that $X = x$, for $x = 1, 2, \dots, 6$, given that A occurs?

What is $\mathbb{P}(X = 1 | A)$?

Start with events. The event A can be written as $A = \{\text{the value of } X \text{ is an even number}\}$. To compute $\mathbb{P}(X = 1 | A)$ we must compute $\mathbb{P}(A)$ (which is $1/2$) and $\mathbb{P}(\{X = 1\} \cap A)$.

$$\{s : X(s) = 1\} \cap A = \{\text{the value of } X \text{ is 1 and the value of } X \text{ is an even number}\} = \emptyset$$

$$\text{since 1 is not an even number. So } \mathbb{P}(X = 1 | A) = \frac{\mathbb{P}(\{X = 1\} \cap A)}{\mathbb{P}(A)} = \frac{0}{1/2} = 0.$$

What is $\mathbb{P}(X = 2 | A)$?

Again, start with events. Now $\{X = 2\} \cap A = \{X = 2\}$ since $\{X = 2\} \subset A$. (In words, as long as we are requiring $X = 2$, the extra requirement that X is even is redundant.) So

$$\mathbb{P}(X = 2 | A) = \frac{\mathbb{P}(\{X = 2\} \cap A)}{\mathbb{P}(A)} = \frac{\mathbb{P}(X = 2)}{\mathbb{P}(X \text{ is even})} = \frac{1/6}{1/2} = \frac{1}{3}.$$

You can do this for all possible values x of X , and you will find that you get 0 if x is odd and $\frac{1}{3}$ if x is even.

This answer makes sense: if you know that the number X on the die is even, then X can only be 2, 4 or 6, and each one is equally likely. The distribution of X given A is then

$$\mathbb{P}(X = x | A) = \begin{cases} \frac{1}{3}, & \text{if } x = 2, 4, 6 \\ 0, & \text{otherwise.} \end{cases}$$

In general we can define the *conditional distribution* of a random variable X given the event A occurs in the following way:

Definition: The conditional probability mass function of X given that an event A occurs, where $\mathbb{P}(A) > 0$, is

$$f_{X|A}(x) = \mathbb{P}(X = x | A) \quad \text{for all } x.$$

Example 7.10 (Example 7.1 continued). Let X, Y have the joint p.m.f. from Example 7.1. What is the conditional distribution of X given that $X + Y$ is even?

With $A = \{X + Y \text{ is even}\}$, we previously computed $\mathbb{P}(A) = 0.3$. We can compute

$$\begin{aligned} \mathbb{P}(\{X = 1\} \cap A) &= \mathbb{P}(X = 1, Y = 3) = 0.2, \\ \mathbb{P}(\{X = 2\} \cap A) &= \mathbb{P}(X = 2, Y = 6) = 0, \\ \mathbb{P}(\{X = 3\} \cap A) &= \mathbb{P}(X = 3, Y = 3) = 0.1 \end{aligned}$$

so

$$f_{X|A}(x) = \begin{cases} \frac{0.2}{0.3} = 2/3 & \text{if } x = 1, \\ \frac{0.1}{0.3} = 1/3 & \text{if } x = 3, \\ 0 & \text{otherwise.} \end{cases}$$

Note that the set of (conditionally) possible values of X changes when we condition on A .

Example 7.11 (Example 7.2 continued). Let X and Y be two independent fair dice rolls, and define $Z = X + Y$. What is the distribution of Z given that $X = 3$?

Set $A = \{X = 3\}$. Then

$$\{Z = z\} \cap A = \{Z = z, X = 3\} = \{X = 3, X + Y = z\} = \{X = 3, Y = z - 3\}.$$

This event has probability $1/36$ if $z - 3 \in \{1, 2, 3, 4, 5, 6\}$ and has probability 0 otherwise. Since $\mathbb{P}(A) = \mathbb{P}(X = 3) = 1/6$,

$$f_{Z|A}(z) = \begin{cases} \frac{1/36}{1/6} = \frac{1}{6} & \text{if } z = 4, 5, 6, 7, 8, 9, \\ 0 & \text{otherwise.} \end{cases}$$

This makes sense: even when we condition on $X = 3$, Y is equally likely to take any of the values $1, 2, 3, 4, 5, 6$ (since Y is independent of X) and on the other hand we can rewrite the formula for Z as $Z = Y + 3$.

We can also define the conditional distribution of a random variable X conditional on an event $\{Y = y\}$ such that $\mathbb{P}(Y = y) > 0$, since $\{Y = y\} = \{s: Y(s) = y\}$ is just a special type of event.

Definition: The conditional probability mass function of X given Y is

$$f_{X|Y}(x|y) = \mathbb{P}(X = x | Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{f_{X,Y}(x, y)}{f_Y(y)} \quad \text{for all } x,$$

for any y such that $f_Y(y) > 0$.

Example 7.12 (Example 7.4 continued). Draw two cards from a standard deck. Set X to be 1 if the first card is an Ace, and set X to be 0 otherwise. Set Y to be 1 if the second card is an Ace, and set Y to be 0 otherwise. Compute the conditional p.m.f. of X given Y .

$$\begin{aligned} f_{X|Y}(0|0) &= \mathbb{P}(X = 0 | Y = 0) = \frac{f_{X,Y}(0, 0)}{f_Y(0)} = \frac{48 \cdot 47/52 \cdot 51}{48/52} = \frac{47}{51}, \\ f_{X|Y}(1|0) &= \mathbb{P}(X = 1 | Y = 0) = \frac{f_{X,Y}(1, 0)}{f_Y(0)} = \frac{4 \cdot 48/52 \cdot 51}{48/52} = \frac{4}{51}, \\ f_{X|Y}(0|1) &= \mathbb{P}(X = 0 | Y = 1) = \frac{f_{X,Y}(0, 1)}{f_Y(1)} = \frac{48 \cdot 4/52 \cdot 51}{4/52} = \frac{48}{51}, \\ f_{X|Y}(1|1) &= \mathbb{P}(X = 1 | Y = 1) = \frac{f_{X,Y}(1, 1)}{f_Y(1)} = \frac{4 \cdot 3/52 \cdot 51}{4/52} = \frac{3}{51}. \end{aligned}$$

For $y \in \{0, 1\}$ and any other x , we obtain $f_{X|Y}(x|y) = 0/f_Y(y) = 0$.

For $y \notin \{0, 1\}$, the denominator is $f_Y(y) = 0$ and $f_{X|Y}(x|y)$ is undefined. So

$$f_{X|Y}(x|y) = \begin{cases} 47/51 & \text{if } x = 0, y = 0, \\ 4/51 & \text{if } x = 1, y = 0, \\ 48/51 & \text{if } x = 0, y = 1, \\ 3/51 & \text{if } x = 1, y = 1, \\ 0 & \text{for } y \in \{0, 1\} \text{ and other values of } x. \end{cases}$$

We can summarise these results in a table:

$f_{X Y}(x y)$		y		
		0	1	
x	0	$\frac{47}{51}$	$\frac{48}{51}$	*
	1	$\frac{4}{51}$	$\frac{3}{51}$	*
		1	1	*

This is a different kind of table than for the joint p.m.f.'s, so it is important to label it as describing $f_{X|Y}(x|y)$. Summing over a fixed y now gives 1 (for any value of y that is possible). However, summing over a fixed x does not give 1, and the sums in positions marked * are not meaningful.

Sometimes it is possible to find conditional distributions directly. In this case, we can use the reverse formula $f_{X,Y}(x,y) = f_X(x)f_{Y|X}(y|x)$ to find the joint p.m.f. (Note that we used $f_{Y|X}(y|x)$, with X and Y swapped, rather than $f_{X|Y}(x|y)$.)

Example 7.13. Deal two hands of 6 cards each from a standard deck. Let X be the number of Aces in the first hand, and let Y be the number of Aces in the second hand. What are

- (a) the marginal p.m.f. $f_X(x)$,
- (b) the conditional p.m.f. $f_{Y|X}(y|x)$, and
- (c) the joint p.m.f. $f_{X,Y}(x,y)$?

- (a) To produce X , we draw 6 cards without replacement from a total of 52. We count the number of successes, where “success” is defined to mean “Ace”; there are 4 such cards. Hence X has the Hypergeometric distribution with parameters $n = 6$, $M = 4$, and $N = 52$, and the p.m.f. is

$$f_X(x) = \frac{\binom{4}{x} \binom{48}{6-x}}{\binom{52}{6}} \quad \text{for } x = 0, 1, 2, 3, 4,$$

with $f_X(x) = 0$ otherwise.

- (b) We only consider the possible values of X , namely $x = 0, 1, 2, 3, 4$. The event $\{X = x\}$ means that we drew x Aces in the first hand (and therefore $6 - x$ non-Aces). So there are $4 - x$ Aces (and $48 - (6 - x) = 42 + x$ non-Aces) remaining among the 46 cards left in the deck. This fully describes everything we know based on the event $\{X = x\}$.

To produce Y , we now draw 6 cards without replacement from these remaining cards. Other than affecting how many of each kind of cards are left, conditioning on $\{X = x\}$ does not affect this subsequent draw of 6 cards.

[That is, we are still equally likely to draw any remaining card at each step.]

So, conditional on $\{X = x\}$, Y has the Hypergeometric distribution with parameters $n = 6$, $M = 4 - x$, and $N = 46$ and

$$f_{Y|X}(y|x) = \frac{\binom{4-x}{y} \binom{42+x}{6-y}}{\binom{46}{6}} \quad \text{for } y = 0, 1, \dots, 4 - x,$$

with $f_{Y|X}(y|x) = 0$ for other values of y .

- (c)

$$f_{X,Y}(x,y) = f_X(x)f_{Y|X}(y|x) = \frac{\binom{4}{x} \binom{48}{6-x} \binom{4-x}{y} \binom{42+x}{6-y}}{\binom{52}{6} \binom{46}{6}}.$$

[Extension: We could have obtained another formula by conditioning on Y instead of X . What is that formula? Is it the same as the formula given above?]

7.4 Testing for independence

We often need to determine whether two (or more) random variables are independent, preferably without making extensive and difficult calculations. There are several kinds of ways to do this:

- Independence may be *given*.
 - The joint distribution of several variables can be specified by giving their marginal distributions and requiring them to be independent.
- Independence may be *assumed* as part of the model of a real phenomenon.
 - The values of successive dice rolls, coin flips, and other *physically independent* random variables are always assumed to be independent.
 - Sometimes it is necessary to assume that random variables are independent as a simplifying assumption, even when we have no specific reason for thinking this. (For instance, numbers of customers in a shop on successive days.) The real-world relevance of the resulting probabilistic model will depend, to some extent, on how reasonable this assumption is.
 - Random-number generators on a computer (which are actually completely deterministic) produce sequences of numbers that are quite reasonable approximations of independent random numbers.
- Independence may be *known* from general theory.
 - Random variables that depend on disjoint sets of independent random variables are automatically independent; see Theorem 7.2 on page 126 and the notes following it.
 - Random variables that depend on overlapping sets of random variables would, by default, be expected to be dependent. When they are actually independent, this interesting and surprising result usually arises from some hidden structure within the problem.
- Independence may need to be *checked* directly.

The last case is the most challenging and interesting. Although there is no general technique, certain common strategies apply in many situations. These strategies differ depending on whether the random variables are suspected of being independent or dependent:

- *Even one counterexample* to the statement $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ (i.e., even one pair of values x, y for which the equation is false) is enough to prove that X and Y are *dependent*.
- The statement $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ must be proved for *every possible* pair of values x, y to prove that X and Y are *independent*.
Except when the number of possible pairs is small, a general argument is needed.
- In both cases, it is enough to consider x, y that are possible values of X, Y (i.e., for which $f_X(x)$ and $f_Y(y)$ are non-zero).

Example 7.14. Let X and Y have joint p.m.f. given by

$f_{X,Y}(x, y)$		y	
		3	4
x	1	$\frac{1}{4}$	$\frac{1}{12}$
	2	$\frac{1}{2}$	$\frac{1}{6}$
		1	

Are X and Y independent or dependent?

There are only 4 possible pairs of values, and no clear pattern. We compute the marginals by hand and then check $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ for each possible pair of values. This working can be summarised in the table:

$f_{X,Y}(x, y)$		y		
		3	4	
x	1	$\frac{1}{4} = \frac{3}{4} \cdot \frac{1}{3}$	$\frac{1}{12} = \frac{1}{4} \cdot \frac{1}{3}$	$\frac{1}{3}$
	2	$\frac{1}{2} = \frac{3}{4} \cdot \frac{2}{3}$	$\frac{1}{6} = \frac{1}{4} \cdot \frac{2}{3}$	$\frac{2}{3}$
		$\frac{3}{4}$	$\frac{1}{4}$	1

This shows that X and Y are *independent*.

Example 7.15. Three squares are given to you: a 1×1 square, a 2×2 square, and a 3×3 square. Choose one uniformly at random and let A and L be the area and the side length, respectively, of the chosen square. Are A and L independent or dependent?

The random variables A and L are related by $A = L^2$. This relation between A and L strongly suggests that they are dependent, and to confirm this we seek a pair of values a, ℓ that are individually possible but fail to satisfy the relation $a = \ell^2$. For instance, $a = 9$ and $\ell = 2$.

Both events $\{A = 9\}$ and $\{L = 2\}$ have positive probability (corresponding to choosing the third and second squares, respectively).

[That is, $f_A(9)$ and $f_L(2)$ are both positive.]

But the event $\{A = 9, L = 2\}$ is impossible since the relation $A = L^2$ is violated.

[That is, $f_{A,L}(9, 2) = 0$.]

Therefore $f_{A,L}(9, 2) \neq f_A(9)f_L(2)$ and A, L are *dependent*.

Example 7.16. Choose an integer uniformly at random between 1 and 100. Let X be the ones digit and let Y be the tens digit of the number chosen. Are X and Y independent?

[If the number chosen is 9 or less, then $Y = 0$. If the number chosen is 100, then $X = Y = 0$.]

X and Y each have 10 possible values, namely $0, 1, 2, \dots, 9$. For each of the $10^2 = 100$ possible pairs of values $x, y \in \{0, 1, \dots, 9\}$, there is exactly one integer between 1 and 100 that, if chosen, results in $X = x, Y = y$. (The only doubtful case is $x = y = 0$, in which case there is still exactly one corresponding integer, 100.) Since each such integer is equally likely to be chosen, the joint p.m.f. is

$$f_{X,Y}(x, y) = \begin{cases} \frac{1}{100} & \text{if } x, y \in \{0, 1, \dots, 9\}, \\ 0 & \text{otherwise.} \end{cases}$$

For each fixed $x \in \{0, 1, \dots, 9\}$, there are exactly 10 values y for which $f_{X,Y}(x, y) = \frac{1}{100}$, and $f_{X,Y}(x, y) = 0$ otherwise. Adding,

$$f_X(x) = \sum_y f_{X,Y}(x, y) = 10 \cdot \frac{1}{100} = \frac{1}{10} \quad \text{and similarly} \quad f_Y(y) = \frac{1}{10}$$

for $x, y \in \{0, 1, \dots, 9\}$. For such x, y , we have $f_{X,Y}(x, y) = \frac{1}{100} = \frac{1}{10} \cdot \frac{1}{10} = f_X(x)f_Y(y)$. This shows that X and Y are *independent*.

[It is not necessary to check $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ for other values of x, y .]

[This argument must break down if the range of numbers is changed to, say, 1 to 101. It is worth checking precisely which parts of the argument fail.]

Example 7.17 (Example 7.13 continued). Deal two hands of 6 cards each from a standard deck. Let X be the number of Aces in the first hand, and let Y be the number of Aces in the second hand. Are X and Y independent or dependent?

The possible values of X are 0, 1, 2, 3, 4, and the same for Y , since either hand could contain all 4 aces.

[That is, $f_X(x)$ and $f_Y(y)$ are both positive for $x, y = 0, 1, 2, 3, 4$.]

We consider whether these values can occur simultaneously, paying particular attention to the largest or smallest possible values. In this case the largest values $X = 4$ and $Y = 4$ cannot occur simultaneously since there are only 4 possible Aces in total.

[That is, $f_{X,Y}(4, 4) = 0$.]

So $f_{X,Y}(4, 4) \neq f_X(4)f_Y(4)$ and X, Y are *dependent*.

[It is not necessary to calculate the values $f_X(4), f_Y(4)$, only to argue that they are positive.]

Example 7.18 (Variant of Examples 7.13/7.17). Deal two hands of 6 cards each from a standard deck. Let X' be the number of Spades in the first hand, and let Y' be the number of Spades in the second hand. Are X and Y independent or dependent?

Based on the previous result, it is tempting to guess that X' and Y' are dependent. The possible values of X', Y' are 0, 1, 2, 3, 4, 5, 6. This time, however, any pair of values can occur since there are more than 12 Spades and more than 12 non-Spades.

[That is, $f_{X,Y}(x, y) > 0$ for every pair of values x, y with $f_X(x) > 0, f_Y(y) > 0$.]

Copying the argument of Example 5, the conditional distribution of Y given $\{X = x\}$ is Hypergeometric with parameters $n = 6, M = 13 - x$, and $N = 46$. These conditional distributions are different for different x , so Y and X are *dependent*.

The last step in the argument uses the following result.

Theorem 7.3. Random variables X and Y are independent if and only if $f_{X|Y}(x|y)$ does not depend on y (over those values of y for which $f_Y(y) > 0$), or equivalently if and only if $f_{Y|X}(y|x)$ does not depend on x (over those values of x for which $f_X(x) > 0$).

7.5 Relations between distributions

Theorem 7.4. Let $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\lambda')$, independently. Set $Z = X + Y$. Then

- (a) The (marginal) distribution of Z is $\text{Poisson}(\lambda + \lambda')$.
- (b) The conditional distribution of X given Z is $\text{Binomial}(Z, p)$ where $p = \lambda/(\lambda + \lambda')$.

Proof of (a).

Compute the (marginal) p.m.f. of Z by summing over the ways in which $\{Z = z\}$ could occur:

$$\begin{aligned}
 f_Z(z) &= \mathbb{P}(Z = z) = \mathbb{P}(X + Y = z) \\
 &= \sum_{x=0}^z \mathbb{P}(X = x, Y = z - x) \\
 &= \sum_{x=0}^z \mathbb{P}(X = x) \mathbb{P}(Y = z - x) && \text{(by independence)} \\
 &= \sum_{x=0}^z \frac{\lambda^x e^{-\lambda}}{x!} \frac{(\lambda')^{z-x} e^{-\lambda'}}{(z-x)!} \\
 &= e^{-\lambda-\lambda'} \sum_{x=0}^z \frac{\lambda^x (\lambda')^{z-x}}{x! (z-x)!}.
 \end{aligned}$$

The denominator resembles part of a binomial coefficient. To make this more clear, multiply and divide by $z!$ to get:

$$\begin{aligned}
 f_Z(z) &= e^{-\lambda-\lambda'} \sum_{x=0}^z \frac{1}{z!} \cdot \frac{z!}{x!(z-x)!} \lambda^x (\lambda')^{z-x} \\
 &= \frac{e^{-\lambda-\lambda'}}{z!} \sum_{x=0}^z \binom{z}{x} \lambda^x (\lambda')^{z-x}.
 \end{aligned} \tag{1}$$

We can evaluate the sum using the Binomial Theorem $(a + b)^n = \sum_{i=0}^n \binom{n}{i} a^i b^{n-i}$, giving

$$f_Z(z) = \frac{e^{-\lambda-\lambda'}}{z!} (\lambda + \lambda')^z.$$

This is the p.m.f. for a $\text{Poisson}(\lambda + \lambda')$ random variable.

□

Proof of (b).

To obtain the conditional distribution of X given Z , compute the conditional p.m.f.

$$\begin{aligned}
f_{X|Z}(x|z) &= \frac{f_{X,Z}(x, z)}{f_Z(z)} = \frac{\mathbb{P}(X = x, Z = z)}{f_Z(z)} \\
&= \frac{\mathbb{P}(X = x)\mathbb{P}(Z = z | X = x)}{f_Z(z)} \\
&= \frac{\mathbb{P}(X = x)\mathbb{P}(Y = z - x | X = x)}{f_Z(z)} && \text{(since } Y = Z - x \text{ when } X = x) \\
&= \frac{\mathbb{P}(X = x)\mathbb{P}(Y = z - x)}{f_Z(z)} && \text{(by independence)} \\
&= \frac{(\lambda^x e^{-\lambda}/x!)((\lambda')^{z-x} e^{-\lambda'}/(z-x)!)}{(\lambda + \lambda')^z e^{-\lambda-\lambda'}/z!} \\
&= \frac{z!}{x!(z-x)!} \frac{\lambda^x (\lambda')^{z-x}}{(\lambda + \lambda')^z} \\
&= \binom{z}{x} \frac{\lambda^x (\lambda')^{z-x}}{(\lambda + \lambda')^z}
\end{aligned}$$

This conditional p.m.f. looks like a Binomial p.m.f. To make this more clear, split $(\lambda + \lambda')^z$ as $(\lambda + \lambda')^x (\lambda + \lambda')^{z-x}$:

$$\begin{aligned}
f_{X|Z}(x|z) &= \binom{z}{x} \frac{\lambda^x}{(\lambda + \lambda')^x} \frac{(\lambda')^{z-x}}{(\lambda + \lambda')^{z-x}} \\
&= \binom{z}{x} \left(\frac{\lambda}{\lambda + \lambda'} \right)^x \left(\frac{\lambda'}{\lambda + \lambda'} \right)^{z-x}.
\end{aligned}$$

Since $\frac{\lambda'}{\lambda + \lambda'} = 1 - \frac{\lambda}{\lambda + \lambda'}$, we recognise this conditional p.m.f. as the Binomial(z, p) p.m.f. with $p = \lambda/(\lambda + \lambda')$, as claimed.

□

The conditional p.m.f. of X given Z encodes the relative contributions of the probabilities $\mathbb{P}(X = x, Z = z)$ to the marginal p.m.f. $\mathbb{P}(Z = z)$. In fact, apart from the factor $(\lambda + \lambda')^z$ that does not depend on x , the conditional p.m.f. is the same as the summand in the computation of $f_Z(z)$.

Alternatively, we can interpret $\mathbb{P}(X = x | Z = z)$ as the original probability $\mathbb{P}(X = x)$ *weighted* proportionally to the conditional probability of $\{Z = z\}$:

$$f_{X|Z}(x|z) = \frac{f_X(x)f_{Z|X}(z|x)}{f_Z(z)} = \frac{f_X(x)f_{Z|X}(z|x)}{\sum_{x'} f_X(x')f_{Z|X}(z|x')}.$$

This is *Bayes' Theorem* for discrete random variables.

Theorem 7.4 showed that adding two independent Poisson random variables produced another Poisson random variable. Certain other named distributions also have this property:

Theorem 7.5. Let $X \sim \text{Binomial}(m, p)$ and let $Y \sim \text{Binomial}(n, p)$, independently. Set $Z = X + Y$. Then

- (a) The (marginal) distribution of Z is $\text{Binomial}(m + n, p)$.
- (b) The conditional distribution of X given Z is $\text{Hypergeometric}(Z, m, m + n)$.

We will prove part (a) only. In the proof of Theorem 7.4(a), we computed the marginal p.m.f. explicitly. This time we will represent the random variables in terms of simpler random variables to avoid messy algebra.

Proof of (a).

By definition, Binomial random variables count successes in sequences of Bernoulli trials. Let $T_1, T_2, \dots, T_m, T_{m+1}, \dots, T_{m+n}$ be independent Bernoulli random variables with success probability p , i.e., $\mathbb{P}(T_i = 1) = p$, $\mathbb{P}(T_i = 0) = 1 - p$. Then we can represent X and Y as

$$X = T_1 + \dots + T_m, \quad Y = T_{m+1} + \dots + T_{m+n}.$$

[More precisely, if we define new random variables $X' = T_1 + \dots + T_m, Y' = T_{m+1} + \dots + T_{m+n}$, then X', Y' have the same joint distribution as X, Y . Everything in Theorem 7.5 depends only on the joint distribution of X, Y , so we can replace X, Y by X', Y' and proceed.]

Since X and Y depend on non-overlapping sets of independent random variables, X and Y are independent by Theorem 7.2. Then

$$Z = X + Y = T_1 + \dots + T_{m+n},$$

so Z counts the number of successes in $m+n$ Bernoulli trials with success probability p . In particular, Z has the $\text{Binomial}(m + n, p)$ distribution.

□

To interpret the conditional distribution of X given Z :

The event $\{Z = z\}$ means that exactly z out of $n + m$ Bernoulli trials were successes, but we do not know which ones. Since the success probability is constant for all trials, any set of z specified trials has the same probability of being the set of successful trials. So, conditional on $\{Z = z\}$, we could decide which trials were the successful ones by choosing z trials, uniformly at random and without replacement, out of $m + n$ choices. In this equivalent experiment, X counts the number of times we selected a trial numbered 1 to m . So the conditional distribution of X will be Hypergeometric.

Theorem 7.6. Let $\lambda > 0$ and let $0 < p < 1$. Let $N \sim \text{Poisson}(\lambda)$ and, conditional on N , let X have the $\text{Binomial}(N, p)$ distribution. Set $Y = N - X$.

- (a) The (marginal) distribution of X is $\text{Poisson}(\lambda p)$.
- (b) The (marginal) distribution of Y is $\text{Poisson}(\lambda(1 - p))$.
- (c) X and Y are independent.

We can think of a Poisson random variable as the number of arrivals (of customers, raindrops, radioactive decays, etc.) over a given time period, where arrivals happen independently from one moment to another. Theorem 7.6 says that if these each arrival is randomly assigned to either X or Y , all independently, then the number of arrivals of each type will be independent. This is a key feature of Poisson random variables, and in fact property ((c)) can *only* hold when N has a Poisson distribution.

Proof.

We will find the joint p.m.f. of X, Y . Since we are given information about the joint distribution of X and N , we will rewrite $\mathbb{P}(X = x, Y = y)$ in terms of X and N . The event $\{X = x, Y = y\}$ is the same as $\{X = x, N = x + y\}$: if we know X , then the value of N can be inferred from the value of Y and vice versa. So

$$\begin{aligned} f_{X,Y}(x, y) &= \mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x, N = x + y) \\ &= \mathbb{P}(N = x + y) \mathbb{P}(X = x | N = x + y) \\ &= \frac{\lambda^{x+y} e^{-\lambda}}{(x+y)!} \cdot \binom{x+y}{x} p^x (1-p)^{(x+y)-x} \end{aligned}$$

(using the formula $\mathbb{P}(\text{Binomial}(n, p) = x) = \binom{n}{x} p^x (1-p)^{n-x}$ with $n = x + y$)

$$\begin{aligned} &= \frac{\lambda^{x+y} e^{-\lambda}}{(x+y)!} \cdot \frac{(x+y)!}{x!(x+y-x)!} p^x (1-p)^y \\ &= \frac{\lambda^{x+y} p^x (1-p)^y e^{-\lambda}}{x!y!} \end{aligned}$$

To show that X and Y are independent, it is enough to show that $f_{X,Y}(x, y)$ factors as a function of x only times a function of y only. Since $\lambda^{x+y} = \lambda^x \lambda^y$, we can do this:

$$\begin{aligned} f_{X,Y}(x, y) &= e^{-\lambda} \frac{\lambda^x p^x}{x!} \cdot \frac{\lambda^y (1-p)^y}{y!} \\ &= \left(e^{-\lambda p} \frac{(\lambda p)^x}{x!} \right) \left(e^{-\lambda(1-p)} \frac{(\lambda(1-p))^y}{y!} \right) \end{aligned}$$

Thus the joint p.m.f. is the product of two marginal p.m.f.'s, which are the p.m.f.'s for the $\text{Poisson}(\lambda p)$ and $\text{Poisson}(\lambda(1 - p))$ distributions, respectively. This automatically shows that X and Y are independent, and the two factors are the two marginal p.m.f.'s. This proves (a), (b) and (c) all at once.

□

Theorem 7.7 (Other relations between distributions).

1. Let $X \sim \text{Binomial}(N, p)$. Conditional on X , let $Y \sim \text{Hypergeometric}(n, X, N)$, and set $Z = X - Y$. Then
 - (a) The marginal distributions are $Y \sim \text{Binomial}(n, p)$ and $Z \sim \text{Binomial}(N - n, p)$.
 - (b) Y and Z are independent.
2. Let $X, Y \sim \text{Geometric}(p)$, independently. Set $Z = X + Y$. Then
 - (a) The marginal distribution is $Z \sim \text{NegBin}(2, p)$.
 - (b) The conditional distribution of $X + 1$ given Z is $\text{DU}(Z + 1)$.
3. Let $X \sim \text{NegBin}(m, p)$ and $Y \sim \text{NegBin}(n, p)$, independently. Set $Z = X + Y$. Then $Z \sim \text{NegBin}(m + n, p)$.
4. Let $X \sim \text{Hypergeometric}(n, M, N)$. Conditional on X , let $Y \sim \text{Hypergeometric}(X, M', M)$, where $M' \leq M \leq N$. Then $Y \sim \text{Hypergeometric}(n, M', N)$.
5. Let $X \sim \text{Geometric}(p)$. Conditional on X , let $Y \sim \text{NegBin}(X + 1, q)$. Set $Z = X + Y$. Then
 - (a) The marginal distribution of Y is $\text{Geometric}(pq/(1 - q + pq))$
 - (b) The conditional distribution of X given Y is $\text{NegBin}(Y + 1, 1 - q + pq)$.
 - (c) The marginal distribution of Z is $\text{Geometric}(pq)$.
 - (d) The conditional distribution of X given Z is $\text{Binomial}(Z, (1 - p)q/(1 - pq))$
 - (e) The conditional distribution of Y given Z is $\text{Binomial}(Z, (1 - q)/(1 - pq))$.

7.6 Exercises

7.6.1 An experiment involves rolling two fair dice. The first one is 4-sided and the second one is 6-sided, so there are 24 possible outcomes. Let X_1 denote the number showing on the first die, and Y denote the sum of the numbers showing on the two dice.

1. One of the possible outcomes for the experiment is $(1, 2)$. Find $X_1((1, 2))$ and $Y((1, 2))$.
2. Give the probability distribution for X_1 .
3. Find $\mathbb{E}[X_1]$ and $\text{Var}(X_1)$.
4. Find $\mathbb{E}[Y]$.

7.6.2 The definition of mutual independence for three events A_1, A_2, A_3 is:

$$\mathbb{P}(A_1 \cap A_2 \cap A_3) = \mathbb{P}(A_1)\mathbb{P}(A_2)\mathbb{P}(A_3) \quad \text{AND} \quad \begin{aligned} \mathbb{P}(A_1 \cap A_2) &= \mathbb{P}(A_1)\mathbb{P}(A_2) \\ \mathbb{P}(A_2 \cap A_3) &= \mathbb{P}(A_2)\mathbb{P}(A_3) \\ \mathbb{P}(A_1 \cap A_3) &= \mathbb{P}(A_1)\mathbb{P}(A_3) \end{aligned}$$

However, the definition of independence for three random variables only requires $f_{X,Y,Z}(x, y, z) = f_X(x)f_Y(y)f_Z(z)$. Why do we not need to add the condition $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ (and similarly for $f_{X,Z}$ and $f_{Y,Z}$)?

7.6.3 Let $X \sim \text{DU}(4)$. Conditional on X , let $Y \sim \text{DU}(X)$ (i.e., conditional on $\{X = x\}$, let $Y \sim \text{DU}(x)$, for $x = 1, 2, 3, 4$).

1. Find the joint p.m.f. $f_{X,Y}(x, y)$ and the marginal p.m.f. $f_Y(y)$.
2. Find the conditional p.m.f. $f_{X|\{Y=1\}}(y)$ (i.e., the conditional p.m.f. $f_{X|Y}(x|y)$ with $y = 1$).

8 Covariance and conditional expectation

By the end of this chapter you should be able to:

- compute covariances and correlations
- compute conditional expectations
- use conditional expectation to compute expectations involving two random variables

8.1 Covariance

Definition: The *covariance* $\text{Cov}(X, Y)$ of two random variables X and Y is

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

The covariance reflects, in a certain averaged way, to what degree obtaining an above-average value of X is associated with obtaining an above-average value of Y . When $Y = X$, the covariance $\text{Cov}(X, X)$ is the same as the variance $\text{Var}(X)$. When $Y = -X$, the covariance $\text{Cov}(X, -X)$ is $-\text{Var}(X)$; the fact that this covariance is negative reflects the simple observation that above-average values of X are associated with below-average values of $-X$.

Just as with variance, it is easier to compute covariances from the shortcut formula

$$\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

Example 8.1 (Example 7.1 continued). Let X, Y have the joint p.m.f. from Example 7.1. Compute $\text{Cov}(X, Y)$.

Use the shortcut formula $\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$, where we have already computed $\mathbb{E}(XY) = 8.1$. We can compute $\mathbb{E}(X)$ and $\mathbb{E}(Y)$ using the marginal distributions from Example 7.1:

$$\begin{aligned}\mathbb{E}(X) &= (1)(0.6) + (2)(0.1) + (3)(0.3) = 1.7, \\ \mathbb{E}(Y) &= (3)(0.4) + (6)(0.6) = 4.8, \\ \text{Cov}(X, Y) &= 8.1 - (1.7)(4.8) = -0.06.\end{aligned}$$

Example 8.2 (Example 7.2 continued). Let X and Y be two independent fair dice rolls. Compute $\text{Cov}(X, Y)$.

Use the shortcut formula $\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$. Since X and Y have the $\text{DU}(6)$ distribution, $\mathbb{E}(X) = \mathbb{E}(Y) = \frac{7}{2}$. By Theorem 7.2, $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y) = \frac{49}{4}$. So

$$\text{Cov}(X, Y) = \frac{49}{4} - \left(\frac{7}{2}\right)\left(\frac{7}{2}\right) = 0.$$

Example 8.3 (Example 7.4 continued). Draw two cards from a standard deck. Set X to be 1 if the first card is an Ace, and set X to be 0 otherwise. Set Y to be 1 if the second card is an Ace, and set Y to be 0 otherwise. Compute $\text{Cov}(X, Y)$.

$$\mathbb{E}(X) = (0)\left(\frac{48}{52}\right) + (1)\left(\frac{4}{52}\right) = \frac{4}{52} = \frac{1}{13}$$

and similarly $\mathbb{E}(Y) = \frac{1}{13}$ since X and Y have the same marginal distribution. Using the joint p.m.f.,

$$\mathbb{E}(XY) = 0 + 0 + 0 + (1 \cdot 1)\left(\frac{4 \cdot 3}{52 \cdot 51}\right) = \frac{4 \cdot 3}{52 \cdot 51}$$

so

$$\text{Cov}(X, Y) = \frac{4 \cdot 3}{52 \cdot 51} - \frac{4 \cdot 4}{52 \cdot 52} = \frac{4}{52} \left(\frac{3}{51} - \frac{4}{52} \right) = \frac{1}{13} \left(\frac{-4}{13 \cdot 17} \right) = -\frac{4}{2873} = -0.001392 \dots$$

In this example, $\text{Cov}(X, Y)$ is negative, and we say that X and Y are **negatively correlated**. This means that, on average, when X has a larger value (here, this can only be $X = 1$), Y is more likely to have smaller values, and when X has a smaller value, Y is more likely to have a larger value. In words, drawing an Ace as the first card decreases the chance of drawing an Ace as the second card, and not drawing an Ace as the first card increases the chance for the second card.

Theorem 8.1.

(a) For any random variables X, Y ,

$$\begin{aligned}\text{Cov}(X, X) &= \text{Var}(X) \\ \text{Cov}(aX + b, cY + d) &= ac \text{Cov}(X, Y) \\ \text{Cov}(X, Y) &= \text{Cov}(Y, X) \\ \text{Var}(X + Y) &= \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y) \\ \text{Var}(X - Y) &= \text{Var}(X) + \text{Var}(Y) - 2 \text{Cov}(X, Y)\end{aligned}$$

(b) $\text{Cov}(X, Y + Z) = \text{Cov}(X, Y) + \text{Cov}(X, Z)$ and generally

$$\text{Cov}\left(\sum_{i=1}^n X_i, \sum_{j=1}^m Y_j\right) = \sum_{i=1}^n \sum_{j=1}^m \text{Cov}(X_i, Y_j).$$

(c) If X and Y are independent then $\text{Cov}(X, Y) = 0$ and $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$.
Generally, if X_1, \dots, X_n are mutually independent then

$$\text{Var}(X_1 + \dots + X_n) = \text{Var}(X_1) + \dots + \text{Var}(X_n).$$

(d) The reverse is not true: $\text{Cov}(X, Y)$ can be 0 even if X and Y are dependent.

Example 8.4 (Example 7.3 continued). Let X and Y be two random variables such that

$$f_{X,Y}(x, y) = \begin{cases} \frac{1}{5} & \text{if } (x, y) = (1, 0), (0, 1), (-1, 0), (0, -1) \text{ or } (0, 0), \\ 0 & \text{otherwise.} \end{cases}$$

Compute $\text{Cov}(X, Y)$.

$$\mathbb{E}(X) = (-1)f_X(-1) + (0)f_X(0) + (1)f_X(1) = (-1)\left(\frac{1}{5}\right) + (0)\left(\frac{3}{5}\right) + (1)\left(\frac{1}{5}\right) = 0.$$

Similarly $\mathbb{E}(Y) = 0$. For $\mathbb{E}(XY)$, note that the points for which $f_{X,Y}(x, y) \neq 0$ all have $xy = 0$. So $\mathbb{E}(XY) = \sum_{x,y} xy f_{X,Y}(x, y) = 0$ and

$$\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = 0 - 0 \cdot 0 = 0.$$

The covariance is zero and we say that X and Y are *uncorrelated* even though X and Y are dependent. This is an example of Theorem 8.1((d)) above.

Example 8.5. Prove that the variance of a Binomial(n, p) random variable is $np(1 - p)$, as stated in Section 6.2.

Call the Binomial random variable Y . We can express Y as the sum of contributions from n independent Bernoulli trials:

$$Y = X_1 + X_2 + \cdots + X_n$$

where X_1, X_2, \dots, X_n are mutually independent with $\mathbb{P}(X_i = 1) = p$, $\mathbb{P}(X_i = 0) = 1 - p$. By Theorem 8.1((c)),

$$\text{Var}(Y) = \text{Var}(X_1 + \cdots + X_n) = \text{Var}(X_1) + \cdots + \text{Var}(X_n).$$

Each X_i has the same distribution, and therefore the same variance. Since X_i is either 0 or 1, we have $X_i^2 = X_i$, $\mathbb{E}(X_i) = p$ and

$$\text{Var}(X_i) = \mathbb{E}(X_i^2) - \mathbb{E}(X_i)^2 = \mathbb{E}(X_i) - \mathbb{E}(X_i)^2 = p - p^2 = p(1 - p),$$

so

$$\text{Var}(Y) = n \text{Var}(X_i) = np(1 - p)$$

as claimed.

Example 8.6 (Example 7.13 continued). Deal two hands of 6 cards each from a standard deck. Let X be the number of Aces in the first hand, and let Y be the number of Aces in the second hand. Find the covariance of X and Y .

Let Z_i be 1 if the i^{th} card is an Ace, and 0 otherwise, so that $X = Z_1 + \cdots + Z_6$ and $Y = Z_7 + \cdots + Z_{12}$. *[This assumes that all 6 cards in the first hand are dealt before any of the cards in the second hand. If cards are dealt in some other order, just rearrange the Z_i 's. As long as the deck is well-shuffled, this makes no difference to the joint distribution.]*

By Theorem 8.1((b)),

$$\begin{aligned}\text{Cov}(X, Y) &= \text{Cov}\left(\sum_{i=1}^6 Z_i, \sum_{j=7}^{12} Z_j\right) \\ &= \sum_{i=1}^6 \sum_{j=7}^{12} \text{Cov}(Z_i, Z_j)\end{aligned}$$

Note that Z_1, \dots, Z_{12} are *dependent*: for instance, no more than 4 of Z_1, \dots, Z_{12} can have the value 1. However, each pair Z_i, Z_j with $i \neq j$ has the same *joint* distribution, and that joint distribution is the same as the random variables in Example 7.4/8.3. We can therefore use the result of Example 8.3:

$$\text{Cov}(Z_i, Z_j) = -\frac{4}{2873} \quad \text{for each } i \neq j$$

and

$$\text{Cov}(X, Y) = \sum_{i=1}^6 \sum_{j=7}^{12} \left(-\frac{4}{2873}\right) = -\frac{6 \cdot 6 \cdot 4}{2873} = -\frac{144}{2873} = -0.0501218 \dots$$

8.2 Correlation

Covariance values are affected by the magnitudes of X and Y themselves. It is often useful to scale out this effect:

Definition: The *correlation* $\rho_{X,Y}$ of random variables X and Y is

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}.$$

X and Y are called *positively correlated* if $\rho_{X,Y} > 0$, *negatively correlated* if $\rho_{X,Y} < 0$, and *uncorrelated* if $\rho_{X,Y} = 0$.

Theorem 8.2. The correlation satisfies

$$-1 \leq \rho_{X,Y} \leq 1.$$

(Equivalently, $|\text{Cov}(X, Y)| \leq \sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}$.) Moreover $\rho_{X,Y} = \pm 1$ if and only if there are constants $a \neq 0$ and $b \in \mathbb{R}$ such that $\mathbb{P}(Y = aX + b) = 1$, and in that case $\rho_{X,Y} = 1$ if $a > 0$, $\rho_{X,Y} = -1$ if $a < 0$.

Example 8.7 (Example 7.1/8.1 continued). Let X, Y have the joint p.m.f. from Example 7.1. Compute $\rho_{X,Y}$.

We computed $\text{Cov}(X, Y) = -0.06$ in Example 8.1; use the marginal distributions from Example 7.1 to find

$$\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = (1^2)(0.6) + (2^2)(0.1) + (3^2)(0.3) - (1.7)^2 = 0.81,$$

$$\text{Var}(Y) = \mathbb{E}(Y^2) - \mathbb{E}(Y)^2 = (3^2)(0.4) + (6^2)(0.6) - (4.8)^2 = 2.16,$$

$$\rho_{X,Y} = \frac{-0.06}{\sqrt{0.81}\sqrt{2.16}} = -0.04536092$$

to 7 significant figures.

Example 8.8 (Example 8.6 continued). Deal two hands of 6 cards each from a standard deck. Let X be the number of Aces in the first hand, and let Y be the number of Aces in the second hand. Find $\rho_{X,Y}$.

The distribution of X is Hypergeometric($n = 6, M = 4, N = 52$), so we can use the results of Section 6.5 to find the variance:

$$\text{Var}(X) = n \frac{M}{N} \left(1 - \frac{M}{N}\right) \frac{N-n}{N-1} = 6 \frac{4}{52} \left(1 - \frac{4}{52}\right) \frac{52-6}{52-1} = \frac{1104}{2873} = 0.384267.$$

X and Y have the same marginal distributions, so $\text{Var}(Y) = 1104/2873$ also. We calculated $\text{Cov}(X, Y) = -144/2873$ in Example 8.6, so

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}} = \frac{-144/2873}{\left(\sqrt{1104/2873}\right)^2} = -\frac{144}{1104} = -\frac{6}{46} = -0.130435.$$

In this example, X and Y are **negatively correlated**. This reflects the fact that having more Aces in one hand tends to decrease, on average, the number of Aces in the other hand.

8.3 Conditional expectation

Definition: The *conditional expectation* of a random variable X given that an event A occurs, where $\mathbb{P}(A) > 0$, is

$$\mathbb{E}(X | A) = \sum_x x \mathbb{P}(X = x | A) = \sum_x x f_{X|A}(x).$$

The *conditional expectation* of a discrete random variable Y given $\{X = x\}$ is

$$\mathbb{E}(Y | X = x) = \sum_y y \mathbb{P}(Y = y | X = x) = \sum_y y f_{Y|X}(y|x)$$

for those values x such that $f_X(x) > 0$.

Example 8.9 (Example 7.1/7.10 continued). Let X, Y have the joint p.m.f. from Example 7.1. Compute $\mathbb{E}(X | A)$ where $A = \{X + Y \text{ is even}\}$.

We already found the conditional p.m.f. $f_{X|A}(x)$, which is only non-zero for $x = 1, 3$ (see page 128). So

$$\mathbb{E}(X | A) = \sum_x x f_{X|A}(x) = (1)\left(\frac{2}{3}\right) + (3)\left(\frac{1}{3}\right) = \frac{5}{3} = 1.6666\dots$$

Example 8.10 (Example 7.13 continued). Deal two hands of 6 cards each from a standard deck. Let X be the number of Aces in the first hand and let Y be the number of Aces in the second hand. Compute $\mathbb{E}(Y | X = x)$, for $x = 0, 1, 2, 3, 4$.

We previously argued that the conditional distribution of Y given $X = x$ is Hypergeometric with parameters $n = 6$, $M = 4 - x$, and $N = 46$. The mean of such a variable is

$$\frac{nM}{N} = \frac{6(4-x)}{46}.$$

So $\mathbb{E}(Y | X = x) = 6(4 - x)/46$ also.

Generally, if we can identify the conditional distribution, then the conditional expectation is the (ordinary) expectation of that distribution.

Properties of conditional expectation

The conditional expectation operation is the ordinary expectation operation applied to a conditional distribution. In particular, it satisfies many of the familiar properties of expectations:

Theorem 8.3. Let X, Y, Z be random variables, let A be an event with $\mathbb{P}(A) > 0$, let a, b be constants, let $g(x), h(x, y)$ be functions, and let x, z be such that $f_X(x), f_Z(z) > 0$.

(a)

$$\begin{aligned}\mathbb{E}(aX + bY | A) &= a\mathbb{E}(X | A) + b\mathbb{E}(Y | A), \\ \mathbb{E}(aX + bY | Z = z) &= a\mathbb{E}(X | Z = z) + b\mathbb{E}(Y | Z = z).\end{aligned}$$

(b)

$$\begin{aligned}\mathbb{E}(g(X) | A) &= \sum_x g(x) f_{X|A}(x), \\ \mathbb{E}(g(Y) | X = x) &= \sum_y g(y) f_{Y|X}(y|x).\end{aligned}$$

(c)

$$\begin{aligned}\mathbb{E}(XY | X = x) &= \mathbb{E}(xY | X = x) = x\mathbb{E}(Y | X = x), \\ \mathbb{E}(h(X, Y) | X = x) &= \mathbb{E}(h(x, Y) | X = x).\end{aligned}$$

(d) If X and Y are independent then

$$\begin{aligned}\mathbb{E}(Y | X = x) &= \mathbb{E}(Y), & \mathbb{E}(X | Y = y) &= \mathbb{E}(X), \\ \mathbb{E}(g(Y) | X = x) &= \mathbb{E}(g(Y)), & \mathbb{E}(g(X) | Y = y) &= \mathbb{E}(g(X)).\end{aligned}$$

Part (c) says that, when conditioning on $\{X = x\}$, we can use the fact that $X = x$ to simplify the expectation. However, in general $\mathbb{E}(XY | X = x) \neq x\mathbb{E}(Y)$ since conditioning on X changes the (conditional) distribution of Y .

One of the most important uses for conditional expectation is as an intermediate step in computing a complicated expectation.

Theorem 8.4 (Partition Theorem for Expectations, Version 1). For any random variables X and Y ,

$$\mathbb{E}(Y) = \sum_{x: f_X(x) > 0} \mathbb{P}(X = x) \mathbb{E}(Y | X = x)$$

Example 8.11 (Example 7.13/8.10 continued). Deal two hands of 6 cards each from a standard deck. Let X be the number of Aces in the first hand and let Y be the number of Aces in the second hand. Compute $\mathbb{E}(XY)$ and $\text{Cov}(X, Y)$.

Use the result of Example 8.10 as follows:

$$\begin{aligned} \mathbb{E}(XY) &= \sum_x \mathbb{P}(X = x) \mathbb{E}(XY | X = x) && \text{(Partition Theorem for Expectations)} \\ &= \sum_x \mathbb{P}(X = x) \mathbb{E}(xY | X = x) && \text{(Theorem 8.3((c)))} \\ &= \sum_x \mathbb{P}(X = x) x \mathbb{E}(Y | X = x) && \text{(Theorem 8.3((a)) with } a = 0, b = x) \\ &= \sum_x \mathbb{P}(X = x) x (6(4 - x)/46) && \text{(Example 8.10).} \end{aligned}$$

We can recognise the last sum as an expectation involving X only:

$$\mathbb{E}(XY) = \mathbb{E}\left(X (6(4 - X)/46)\right).$$

To compute this expectation, recall that X has the $\text{HYP}(n = 6, M = 4, N = 52)$ distribution and use the formula $\mathbb{E}(X^2) = \text{Var}(X) + (\mathbb{E}(X))^2$ plus the facts from Section 6.5:

$$\begin{aligned} \mathbb{E}(XY) &= \mathbb{E}\left(\frac{6(4X - X^2)}{46}\right) \\ &= \frac{6}{46} (4\mathbb{E}(X) - \mathbb{E}(X^2)) \\ &= \frac{6}{46} \left(4\mathbb{E}(X) - (\text{Var}(X) + (\mathbb{E}(X))^2)\right) \\ &= \frac{6}{46} \left(4 \left(\frac{(6)(4)}{52}\right) - (6) \frac{4}{52} \left(1 - \frac{4}{52}\right) \frac{52 - 6}{52 - 1} - \left(\frac{(6)(4)}{52}\right)^2\right) \end{aligned}$$

which works out to

$$\mathbb{E}(XY) = \frac{36}{221} = 0.162896.$$

Finally, since X and Y are both $\text{Hypergeometric}(n = 6, M = 4, N = 52)$ with $\mathbb{E}(X) = \mathbb{E}(Y) = 6 \cdot 4/52 = 6/13$,

$$\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \frac{36}{221} - \frac{6}{13} \cdot \frac{6}{13} = -\frac{144}{2873},$$

in agreement with our earlier calculation in Example 8.6.

8.4 Prediction

When Y is a discrete random variable, we have studied its probability mass function $f_Y(y)$. The p.m.f. completely describes the distribution of Y . We can think of the p.m.f. as a vector of probability values, one for each possible value of Y . However, this is a fairly complicated object, and it may be difficult or impossible to find.

Often, we want to give a more compact summary of the values Y might take. That is, we want to find a number y_{pred} that *predicts* the value of Y , or gives an indication of *typical values* of Y . There are at least three common numbers that aim to do this:

- the *mode* – the most likely value of Y ; ²
- the *median* – a value y such that $\mathbb{P}(Y \leq y) \geq 1/2$ and $\mathbb{P}(Y \geq y) \geq 1/2$; ³ or
- the *expected value* – $\mathbb{E}(Y) = \sum_y y\mathbb{P}(Y = y)$.

In this course and elsewhere, we most often deal with the expected value. ⁴ That is, if we are asked to predict the value of the random variable Y , we use as our prediction the (non-random) number $y_{\text{pred}} = \mathbb{E}(Y)$.

However, we can do better. Often, we have access to an *explanatory* random variable X , which may be easier to measure than Y . Then we can make a more sophisticated prediction $Y_{\text{pred}} = g(X)$, where we are allowed to use the value of the random variable X . The exact function g will depend on the distribution of Y (and in fact the joint distribution of X and Y) but not on the value of Y . That is, our prediction is a *random variable* Y_{pred} , with the constraint that our prediction is allowed to use the (simple, easy-to-measure) random variable X but not the (complicated, hard-to-measure) random variable Y .

To work out how to make this kind of prediction, we need a way of measuring the “goodness” of a prediction y for a random variable Y . One obvious way is to measure the probability that the predicted value is actually obtained, $\mathbb{P}(Y = y)$. According to this measure, we should choose as our prediction the value y_{pred} that maximises $\mathbb{P}(Y = y)$. This is exactly the definition of the mode. The idea of looking at the mode follows directly from looking at a specific measure of goodness of prediction and then maximising.

In general, it is more natural to measure the “badness”, or error, of a prediction and then minimise. For instance, we could measure the error as $\mathbb{P}(Y \neq y) = 1 - \mathbb{P}(Y = y)$. Then minimising leads to the mode, as before, since $\mathbb{P}(Y \neq y)$ is smallest when $\mathbb{P}(Y = y)$ is largest.

Another possible measure of error is $\mathbb{E}(|Y - y|)$, the expected absolute difference between the actual value and the prediction. It turns out that the minimising value y_{pred} is the median, as defined above. However, the minimiser may not be unique, and for this and other reasons, we rarely use $\mathbb{E}(|Y - y|)$ as a measure of error.

²Note: there could be more than one such value, so the mode may not be unique.

³For discrete random variables, the median may not be unique as defined here. Sometimes (especially for a sample median) a slightly different definition is used.

⁴We have also looked at the variance and standard deviation, but these are measures of the *spread* of the distribution of Y , rather than of actual values of Y .

In most cases, the best way to measure error turns out to be the *mean squared error of prediction* (MSEP):

$$\mathbb{E}((Y - y)^2).$$

Theorem 8.5. Over all predictions of Y by a (non-random) number $y_{\text{pred}} \in \mathbb{R}$, the mean squared error of prediction $\mathbb{E}((Y - y_{\text{pred}})^2)$ is minimised when $y_{\text{pred}} = \mathbb{E}(Y)$.

Proof. Expand the square:

$$\mathbb{E}((Y - y_{\text{pred}})^2) = \mathbb{E}(Y^2 - 2y_{\text{pred}}Y + y_{\text{pred}}^2) = \mathbb{E}(Y^2) - 2y_{\text{pred}}\mathbb{E}(Y) + y_{\text{pred}}^2$$

since y_{pred} is non-random.

Thinking of $\mathbb{E}((Y - y_{\text{pred}})^2)$ as a function of y_{pred} , we see that it is a quadratic function. Complete the square:

$$\begin{aligned}\mathbb{E}((Y - y_{\text{pred}})^2) &= y_{\text{pred}}^2 - 2y_{\text{pred}}\mathbb{E}(Y) + \mathbb{E}(Y^2) - \mathbb{E}(Y)^2 + \mathbb{E}(Y^2) \\ &= (y_{\text{pred}} - \mathbb{E}(Y))^2 + \mathbb{E}(Y^2) - \mathbb{E}(Y)^2 \\ &= (y_{\text{pred}} - \mathbb{E}(Y))^2 + \text{Var}(Y).\end{aligned}$$

This quadratic function is minimised when $y_{\text{pred}} - \mathbb{E}(Y) = 0$, i.e., by the value $y_{\text{pred}} = \mathbb{E}(Y)$. \square

As a by-product of the proof, we obtain another interpretation of the variance:

The variance $\text{Var}(Y)$ is the smallest possible mean squared error of prediction when we predict the random variable Y by a constant.

The prediction $y_{\text{pred}} = \mathbb{E}(Y)$ can be made “in advance,” without using any information from the random experiment. We can equally well define the MSEP for a prediction that does use information from the random experiment: The MSEP for predicting Y by Y_{pred} is

$$\mathbb{E}((Y - Y_{\text{pred}})^2)$$

The simplest kind of prediction is a *linear prediction* $Y_{\text{pred}} = aX + b$. To predict in the best possible way, we want to choose the constants a and b to minimise the MSEP.

Theorem 8.6. Over all linear predictions for Y of the form $Y_{\text{pred}} = aX + b$, the mean squared error of prediction $\mathbb{E}((Y - Y_{\text{pred}})^2)$ is minimised when

$$a = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}, \quad b = \mathbb{E}(Y) - a\mathbb{E}(X)$$

Theorem 8.6 says that the best linear prediction for Y is

$$Y_{\text{pred}} = \mathbb{E}(Y) + \frac{\text{Cov}(X, Y)}{\text{Var}(X)}(X - \mathbb{E}(X)).$$

This relation between Y_{pred} and X is also called the *least squares regression line*. The first term is the constant prediction $y_{\text{pred}} = \mathbb{E}(Y)$ from Theorem 8.5. We have made our prediction random by adding a suitably chosen multiple of the recentred explanatory random variable $X - \mathbb{E}(X)$. We also obtain an interpretation of the covariance:

The covariance $\text{Cov}(X, Y)$ indicates (after dividing by $\text{Var}(X)$) the appropriate multiple of $(X - \mathbb{E}(X))$ in order to obtain the best linear prediction for Y .

In other words,

The covariance $\text{Cov}(X, Y)$ indicates (after dividing by $\text{Var}(X)$) the slope of the least squares regression line for Y versus X .

Proof of Theorem 8.6.

We want to minimise the MSEP $\mathbb{E}((Y - Y_{\text{pred}})^2) = \mathbb{E}((Y - (aX + b))^2)$ over all possible choices of the constants a, b . To start, think of a as fixed and set b to be the best possible value (for that given value of a). Since the value of b does not affect the possible choices for a , we can then continue by finding the best possible value of a (with the best possible value of b inserted in our formula).

For a fixed, we can rewrite the MSEP as

$$\mathbb{E}((Y - (aX + b))^2) = \mathbb{E}((Y - aX - b)^2) = \mathbb{E}(((Y - aX) - b)^2).$$

Then we can apply Theorem 8.5 to see that the minimum occurs when $b = \mathbb{E}(Y - aX) = \mathbb{E}(Y) - a\mathbb{E}(X)$. Henceforth we will assume that b has this value.

Substitute this value of b , then expand and complete the square:

$$\begin{aligned} \mathbb{E}((Y - (aX + b))^2) &= \mathbb{E}((Y - \mathbb{E}(Y) - a(X - \mathbb{E}(X)))^2) \\ &= \mathbb{E}((Y - \mathbb{E}(Y))^2) - 2a\mathbb{E}((Y - \mathbb{E}(Y))(X - \mathbb{E}(X))) + a^2\mathbb{E}((X - \mathbb{E}(X))^2) \\ &= \text{Var}(Y) - 2a\text{Cov}(X, Y) + a^2\text{Var}(X) \\ &= \text{Var}(X) \left(a^2 - 2a\frac{\text{Cov}(X, Y)}{\text{Var}(X)} + \frac{\text{Cov}(X, Y)^2}{\text{Var}(X)^2} - \frac{\text{Cov}(X, Y)^2}{\text{Var}(X)^2} \right) + \text{Var}(Y) \\ &= \text{Var}(X) \left(a - \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \right)^2 + \text{Var}(Y) - \frac{\text{Cov}(X, Y)^2}{\text{Var}(X)} \end{aligned}$$

This quadratic is minimised when $a = \text{Cov}(X, Y)/\text{Var}(X)$, and we already found $b = \mathbb{E}(Y) - a\mathbb{E}(X)$ as noted above. \square

As a by-product of the proof, the smallest possible MSEP (for a linear prediction $Y_{\text{pred}} = aX + b$) is $\text{Var}(Y) - \text{Cov}(X, Y)^2/\text{Var}(X)$. This is at least as small as $\text{Var}(Y)$, which was the smallest possible MSEP when predicting by a constant y_{pred} : see Theorem 8.5. The fact that the minimum value is now smaller reflects the fact that we are now minimising over a wider class of possible predictions.

To see how much smaller, substitute the formula $\text{Cov}(X, Y) = \rho_{X,Y}\sqrt{\text{Var}(X)\text{Var}(Y)}$. Then

$$\min_{a,b \in \mathbb{R}} \mathbb{E}((Y - (aX + b))^2) = \text{Var}(Y) - \frac{(\rho_{X,Y}^2 \text{Var}(X) \text{Var}(Y))}{\text{Var}(X)} = \text{Var}(Y)(1 - \rho_{X,Y}^2).$$

So the correlation $\rho_{X,Y}$ has the following interpretation:

The square of the correlation, $\rho_{X,Y}^2$, measures the fraction by which the mean squared error of prediction decreases when we predict Y by a linear prediction $Y_{\text{pred}} = aX + b$ instead of a constant.

Finally, we could make a prediction $Y_{\text{pred}} = g(X)$ that depends on the explanatory random variable X , but not necessarily in a linear way.

Theorem 8.7. Over all possible predictions for Y of the form $Y_{\text{pred}} = g(X)$, where g is an arbitrary function, the mean squared error of prediction $\mathbb{E}((Y - Y_{\text{pred}})^2)$ is minimised when

$$g(x) = \mathbb{E}(Y \mid X = x) \quad \text{for all } x \text{ with } f_X(x) > 0.$$

Proof. Rewrite the MSEP by using the Partition Theorem for Expectations:

$$\begin{aligned} \mathbb{E}((Y - Y_{\text{pred}})^2) &= \mathbb{E}((Y - g(X))^2) \\ &= \sum_{x: f_X(x) > 0} \mathbb{P}(X = x) \mathbb{E}((Y - g(X))^2 \mid X = x) \quad (\text{Theorem 8.4}) \\ &= \sum_{x: f_X(x) > 0} \mathbb{P}(X = x) \mathbb{E}((Y - g(x))^2 \mid X = x) \quad (\text{Theorem 8.3(c)}) \end{aligned}$$

Since $g(x)$ is non-random, we can expand and complete the square, as in the proof of Theorem 8.5:

$$\begin{aligned} \mathbb{E}((Y - Y_{\text{pred}})^2) &= \sum_{x: f_X(x) > 0} \mathbb{P}(X = x) \mathbb{E}(Y^2 - 2g(x)Y + g(x)^2 \mid X = x) \\ &= \sum_{x: f_X(x) > 0} \mathbb{P}(X = x) [\mathbb{E}(Y^2 \mid X = x) - 2g(x)\mathbb{E}(Y \mid X = x) + g(x)^2] \quad (\text{Theorem 8.3(a)}) \\ &= \sum_{x: f_X(x) > 0} \mathbb{P}(X = x) [(g(x) - \mathbb{E}(Y \mid X = x))^2 + \mathbb{E}(Y^2 \mid X = x) - (\mathbb{E}(Y \mid X))^2] \end{aligned}$$

This is a sum of quadratic functions in the variables $g(x)$, over all x such that $f_X(x) > 0$. We are allowed to choose each value $g(x)$ separately, so the minimum occurs when $g(x) = \mathbb{E}(Y \mid X = x)$ for each such x . (The other values of $g(x)$, i.e., when $f_X(x) = 0$, make no difference to the MSEP.) This is exactly what Theorem 8.7 claims. \square

The prediction Y_{pred} given by Theorem 8.7 is the best possible random prediction (in the sense of MSEP) that depends only on X . In the next section, we will give Y_{pred} a special name and notation, and study its properties.

8.5 Conditional expectation as a random variable

Definition: Given discrete random variables X and Y , write

$$\psi_Y(x) = \mathbb{E}(Y | X = x).$$

The *conditional expectation* $\mathbb{E}(Y | X)$ is defined to be the *random variable*

$$\mathbb{E}(Y | X) = \psi_Y(X).$$

Example 8.12 (Example 7.13/8.10/8.11 continued). Deal two hands of 6 cards each from a standard deck. Let X be the number of Aces in the first hand and let Y be the number of Aces in the second hand. Find $\mathbb{E}(Y | X)$.

We previously found $\mathbb{E}(Y | X = x) = 6(4 - x)/46$ for $x = 0, 1, 2, 3, 4$. Thus $\psi_Y(x) = 6(4 - x)/46$ and

$$\mathbb{E}(Y | X) = \frac{6(4 - X)}{46}.$$

Example 8.13. You have two fair dice, one with 4 sides (numbered 1 to 4) and one with 6 sides (numbered 1 to 6). Roll the 4-sided die, and let X be the number you roll. Then roll the 6-sided die repeatedly until you roll a number higher than your first roll (i.e., higher than X). Let Y be the number of times you roll the second die, not including the last time (i.e., the number of times you roll X or less on the second die before stopping). Find $\mathbb{E}(Y | X)$.

Consider the rolls of the second die as trials, with “success” defined to mean rolling higher than X . Conditional on $\{X = x\}$, the trials have a constant success probability $p = \mathbb{P}(\text{roll more than } x) = (6 - x)/6$ per trial, independently for each roll. Thus, conditional on $\{X = x\}$, Y follows a Geometric distribution with $p = (6 - x)/6$, and

$$\psi_Y(x) = \mathbb{E}(Y | X = x) = \frac{1 - p}{p} = \frac{x}{6 - x}, \quad \mathbb{E}(Y | X) = \psi_Y(X) = \frac{X}{6 - X}.$$

Note that $\mathbb{E}(Y | X)$ is a random variable, not a number. Moreover, $\mathbb{E}(Y | X)$ is a function of X only: the value of $\mathbb{E}(Y | X)$ depends on the value of the random variable X , but not on the value of the random variable Y .

Theorem 8.8 (Partition Theorem for Expectations, Version 2). For any random variables X and Y ,

$$\mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y | X)).$$

Interpret Theorem 8.8 with some care. In the right-hand side, $\mathbb{E}(Y | X)$ is a new random variable whose value depends only on the value of X . Theorem 8.8 says that this random variable has the same expectation as Y .

The two versions of the Partition Theorem for Expectations (Theorem 8.4 and Theorem 8.8) are equivalent. To see this, recall that the values of the random variable $\mathbb{E}(Y | X)$ are the numbers $\psi_Y(x) = \mathbb{E}(Y | X = x)$, and use the formula for the expectation of a function of a random variable, Theorem 5.1.

Example 8.14 (Example 8.12 continued). Deal two hands of 6 cards each from a standard deck. Let X be the number of Aces in the first hand and let Y be the number of Aces in the second hand. Use $\mathbb{E}(Y | X)$ to find $\mathbb{E}(Y)$.

$$\mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y | X)) = \mathbb{E}\left(\frac{6(4 - X)}{46}\right) = \frac{6}{46}(4 - \mathbb{E}(X)).$$

We know that $X \sim \text{Hypergeometric}(n = 6, M = 4, N = 52)$, so $\mathbb{E}(X) = 6 \cdot 4/52$ and

$$\mathbb{E}(Y) = \frac{6}{46}\left(4 - \frac{6 \cdot 4}{52}\right) = \frac{6 \cdot 4}{46}\left(1 - \frac{6}{52}\right) = \frac{6 \cdot 4}{46} \cdot \frac{46}{52} = \frac{6 \cdot 4}{52}.$$

This is not surprising, since X and Y have the same marginal distributions and therefore the same expectations.

Example 8.15 (Example 8.13 continued). You have two fair dice, one with 4 sides (numbered 1 to 4) and one with 6 sides (numbered 1 to 6). Roll the 4-sided die, and let X be the number you roll. Then roll the 6-sided die repeatedly until you roll a number higher than your first roll (i.e., higher than X). Let Y be the number of times you roll the second die, not including the last time (i.e., the number of times you roll X or less on the second die before stopping). Find $\mathbb{E}(Y)$.

By the Partition Theorem for Expectations,

$$\mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y | X)) = \mathbb{E}\left(\frac{X}{6 - X}\right).$$

Since $X \sim \text{DU}(4)$,

$$\mathbb{E}(Y) = \frac{1}{4} \cdot \frac{1}{6 - 1} + \frac{1}{4} \cdot \frac{2}{6 - 2} + \frac{1}{4} \cdot \frac{3}{6 - 3} + \frac{1}{4} \cdot \frac{4}{6 - 4} = \frac{37}{40} = 0.925.$$

Theorem 8.9. Let X, Y, Z be random variables and let g be a function.

(a)

$$\mathbb{E}(aX + bY | Z) = a\mathbb{E}(X | Z) + b\mathbb{E}(Y | Z).$$

(b)

$$\mathbb{E}(g(Y) | X) = \psi'(X),$$

where

$$\psi'(x) = \mathbb{E}(g(Y) | X = x) = \sum_y g(y)f_{Y|X}(y|x).$$

(c)

$$\begin{aligned}\mathbb{E}(XY | X) &= X\mathbb{E}(Y | X), \\ \mathbb{E}(g(X)Y | X) &= g(X)\mathbb{E}(Y | X).\end{aligned}$$

(d)

$$\mathbb{E}(g(X)\mathbb{E}(Y | X)) = \mathbb{E}(g(X)Y).$$

(e) If X and Y are independent then

$$\mathbb{E}(g(Y) | X) = \mathbb{E}(g(Y)), \quad \mathbb{E}(g(X) | Y) = \mathbb{E}(g(X)).$$

Property ((c)) says that, when conditioning on X , we may treat X (or functions of X) as a constant and factor them out of the conditional expectation. This is analogous to Theorem 8.3((c)). Property ((d)) can be obtained from ((c)) by taking expectations of both sides and applying the Partition Theorem for Expectations, Theorem 8.8.

Example 8.16 (Example 7.13/8.10/8.11/8.12 continued). Deal two hands of 6 cards each from a standard deck. Let X be the number of Aces in the first hand and let Y be the number of Aces in the second hand. Use $\mathbb{E}(Y | X)$ and the properties of conditional expectation to compute $\mathbb{E}(XY)$.

Recalling from Example 8.12 that $\mathbb{E}(Y | X) = 6(4 - X)/46$,

$$\begin{aligned}
 \mathbb{E}(XY) &= \mathbb{E}(\mathbb{E}(XY | X)) && \text{(Partition Theorem for Expectations, Version 2)} \\
 &= \mathbb{E}(X\mathbb{E}(Y | X)) && \text{(Theorem 8.9((c)))} \\
 &= \mathbb{E}\left(X \frac{6}{46} (4 - X)\right) && \text{(Example 8.12)} \\
 &= \frac{6}{46} (4\mathbb{E}(X) - \mathbb{E}(X^2))
 \end{aligned}$$

We obtained the same formula in Example 8.11. We can finish the calculation in the same way:

$$\begin{aligned}
 \mathbb{E}(XY) &= \frac{6}{46} \left(4\mathbb{E}(X) - \left(\text{Var}(X) + (\mathbb{E}(X))^2 \right) \right) \\
 &= \frac{6}{46} \left(4 \left(\frac{(6)(4)}{52} \right) - (6) \frac{4}{52} \left(1 - \frac{4}{52} \right) \frac{52 - 6}{52 - 1} - \left(\frac{(6)(4)}{52} \right)^2 \right) \\
 &= \frac{36}{221} = 0.162896.
 \end{aligned}$$

As in Example 8.11, we can use this calculation to find the covariance via $\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = -144/2873$.

This calculation (using $\mathbb{E}(Y | X)$ and Theorem 8.9) closely paralleled the calculation in Example 8.11 (using $\mathbb{E}(Y | X = x)$ and Theorem 8.3). In this calculation, the two methods work equally well. The next two results, however, rely on interpreting conditional expectations as variables.

Theorem 8.10. For any random variables X, Y ,

$$\text{Cov}(X, Y) = \text{Cov}(X, \mathbb{E}(Y | X)).$$

Example 8.17 (Example 7.13/8.10/8.11/8.12/8.16 continued). Deal two hands of 6 cards each from a standard deck. Let X be the number of Aces in the first hand and let Y be the number of Aces in the second hand. Use Theorem 8.10 to compute $\text{Cov}(X, Y)$ and $\rho_{X,Y}$.

Using $\mathbb{E}(Y | X) = 6(4 - X)/46$ from Example 8.12,

$$\begin{aligned} \text{Cov}(X, Y) &= \text{Cov}(X, \mathbb{E}(Y | X)) && \text{(by Theorem 8.10)} \\ &= \text{Cov}\left(X, \frac{6(4 - X)}{46}\right) && \text{(by Example 8.12)} \\ &= \frac{6}{46} \text{Cov}(X, 4 - X) = \frac{6}{46} \text{Cov}(X, -X) && \text{(by Theorem 8.1((a)))} \\ &= -\frac{6}{46} \text{Cov}(X, X) = -\frac{6}{46} \text{Var}(X). \end{aligned}$$

Since X is a HYP($n = 6, M = 4, N = 52$) random variable,

$$\text{Cov}(X, Y) = -\frac{6}{46} \text{Var}(X) = -\frac{6}{46} \left((6) \frac{4}{52} \left(1 - \frac{4}{52} \right) \frac{52 - 6}{52 - 1} \right) = -\frac{144}{2873}.$$

To compute $\rho_{X,Y}$, remember that $\text{Var}(X) = \text{Var}(Y)$ (since X and Y have the same marginal distributions) so

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}} = \frac{\text{Cov}(X, Y)}{(\sqrt{\text{Var}(X)})^2} = \frac{-6 \text{Var}(X)/46}{\text{Var}(X)} = -\frac{6}{46},$$

and we do not even need the value of $\text{Var}(X)$ to make this calculation.

Theorem 8.11 (Variance Partition Formula). For random variables X, Y , define

$$\text{Var}(Y | X) = \mathbb{E}(Y^2 | X) - (\mathbb{E}(Y | X))^2,$$

the conditional variance of Y given X . Then

$$\text{Var}(Y) = \mathbb{E}(\text{Var}(Y | X)) + \text{Var}(\mathbb{E}(Y | X)).$$

8.6 Probability, statistics and data

So far, we have taken the viewpoint of *probability*: start with an explicitly defined model and use the model to *deduce* what would happen if we ran the experiment repeatedly. For instance, computing the probability of an event tells us in what fraction of experiments, on average, we would expect to see the event occur.

The viewpoint of *statistics* is the reverse: we have a model with missing information, but we can run an data-gathering experiment to try to *infer* properties of the unknown model. The data might come in several forms, and we might have varying degrees of understanding about the process that created the data. For instance:

- (a) Poll 1000 people and ask them whether they will vote “yes” or “no” in a referendum.

If we assume the total number of people available is much larger than 1000, we can assume that the yes-or-no responses form a sequence of Bernoulli trials. However, the success probability p – the fraction of all people who would say “yes” if polled – is unknown.

- (b) Roll 10 marbles off a table, and measure how fast each marble rolled and how far from the table each marble landed.

A plausible guess is that the landing distance is proportional to the speed, but, even if this guess is correct, both distances and speeds may be subject to measurement errors of unknown size.

- (c) Look up the ages of all New Zealand prime ministers (at least, all the prime ministers so far) at the time when they became prime minister.

We might imagine that the prime ministers so far are examples of a “typical” New Zealand prime minister representing all possible past and future prime ministers. Apart from some basic conditions – for instance, prime ministers should be old enough to vote – we have no obvious information about the age of a typical New Zealand prime minister.

To use the data effectively, we need to answer two basic and related questions:

- How should we condense the data into a single number, called a *statistic*, that represents some underlying feature, called a *parameter*, of the unknown model?
- What does this statistic tell us about the true parameter?

8.6.1 Choosing a random data point

Consider first a simple situation: we have made repeated measurements and obtained n data values x_1, x_2, \dots, x_n , each of which is a real number. For the moment, we will think of x_1, x_2, \dots, x_n as fixed. Then, for instance, we can compute the *sample mean*

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}.$$

Does the sample mean correspond to the mean (i.e., expectation) of an actual random variable?

The answer is yes. Define \hat{X} to be the result of choosing one of the values x_1, x_2, \dots, x_n uniformly at random. (If some of the values are repeated, this choice includes multiplicity: more precisely, we set

$$\hat{X} = x_N \quad \text{where} \quad N \sim \text{DU}(n)$$

so that an x -value that occurs twice is twice as likely to be chosen as an x -value that is not repeated.) Then

$$\mathbb{E}(\hat{X}) = \bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

by the formula for the expectation of a function of the random variable N .

8.6.2 Data as the values of random variables

The randomly chosen data value \hat{X} makes a rudimentary link between data and probability. To go further, we need to formalise our ideas about the data: we need to specify the mechanics of the underlying *data-generating process* which we are trying to understand.

The fundamental idea linking probability and statistics is:

The data-generating process is *random*.

That is, we should not think of the data as the fixed values but rather as *values of random variables*. Thus, we will replace x_1, x_2, \dots, x_n (real numbers that we thought of as fixed) by X_1, X_2, \dots, X_n (random variables with real-number values). We can still form the *sample mean*

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n},$$

we can still select one of the data values at random

$$\hat{X} = X_N \quad \text{where } N \sim \text{DU}(n) \text{ is independent of } X_1, X_2, \dots, X_n,$$

and the sample mean is still an expectation, but now a conditional expectation given the data:

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} = \mathbb{E}(\hat{X} \mid X_1, \dots, X_n).$$

If we want to think of the data values, X_1, X_2, \dots, X_n , as random variables, we need to describe their joint distribution. The simplest model is to assume that X_1, X_2, \dots, X_n are independent with a common distribution:

$$X_i \sim X \quad \text{with} \quad X_1, X_2, \dots, X_n \text{ independent.}$$

We say that X_1, X_2, \dots, X_n form a *sample* from the *population* for which a typical member is represented by X . The data X_1, X_2, \dots, X_n that we end up seeing – the answers to the poll; the ten marble measurements; the ages of prime ministers – may be interesting on their own, but X_1, X_2, \dots, X_n are primarily interesting as a means of studying the unknown distribution of X .

8.6.3 Estimators, sampling distributions and bias

Very commonly, we want to know the population mean, $\mathbb{E}(X)$ – the expected value over the whole population. Since we do not understand the full complexities of the process that generated our data (i.e., since we do not know the distribution of X), we cannot find $\mathbb{E}(X)$ by a theoretical computation. Instead, we can use *estimate* the unknown value $\mathbb{E}(X)$ using the data only. The sample mean \bar{X} is a function of the data X_1, \dots, X_n only, so we can use \bar{X} as an *estimator* to try to infer $\mathbb{E}(X)$. We want to know how well or how badly \bar{X} does as an estimator of $\mathbb{E}(X)$.

Note: this reverses the perspective of Section 8.4, where we thought about predicting the unknown value of a random variable Y using non-random values such as $\mathbb{E}(Y)$, which we thought of as being “known.”

Since the data X_1, X_2, \dots, X_n are random variables, any statistic we compute from the data, such as the sample mean $\bar{X} = \frac{1}{n}(X_1 + \dots + X_n)$, is also a random variable. In general, \bar{X} and X will have different distributions (except when $n = 1$ or when X is a constant) and the distribution of \bar{X} is called the *sampling distribution*.

Assessing whether \bar{X} is a good estimator of $\mathbb{E}(X)$ means asking whether the (random) values of \bar{X} are close, in a suitable sense, to the (constant) value $\mathbb{E}(X)$. We can start by finding the expectation of \bar{X} :

$$\begin{aligned}\mathbb{E}(\bar{X}) &= \mathbb{E}\left(\frac{X_1 + \dots + X_n}{n}\right) \\ &= \frac{1}{n}\mathbb{E}(X_1 + \dots + X_n) \\ &= \frac{1}{n}(\mathbb{E}(X_1) + \dots + \mathbb{E}(X_n)) \\ &= \mathbb{E}(X)\end{aligned}$$

since X_1, \dots, X_n all have the same marginal distribution as X . Loosely speaking, \bar{X} equals $\mathbb{E}(X)$ on average. More formally, we translate $\mathbb{E}(\bar{X}) = \mathbb{E}(X)$ into the statement that \bar{X} is an *unbiased estimator* of $\mathbb{E}(X)$.

We can further compute

$$\text{Var}(\bar{X}) = \left(\frac{1}{n}\right)^2 \text{Var}(X_1 + \dots + X_n) = \frac{n \text{Var}(X)}{n^2} = \frac{\text{Var}(X)}{n}$$

using the properties of variance from Theorem 8.1. Thus, although \bar{X} and X have the same expected value, the deviations of \bar{X} away from this expected value are smaller by a factor $1/\sqrt{n}$.⁵ In particular, when n is large, \bar{X} is tightly concentrated around the constant value $\mathbb{E}(X)$. Roughly speaking, if the sample size is large enough – and if the required independence assumptions are satisfied – then the sample mean is likely to be very close to the truth.

This fact is generalised in the Strong Law of Large Numbers:

Theorem 8.12. If X_1, X_2, \dots are independent with common distribution $X_i \sim X$, and if $\mathbb{E}(X)$ exists, then

$$\mathbb{P}\left(\lim_{n \rightarrow \infty} \frac{X_1 + \dots + X_n}{n} = \mathbb{E}(X)\right) = 1.$$

⁵In the sense that the standard deviation $\sigma_{\bar{X}} = \sqrt{\text{Var}(\bar{X})}$ is smaller by a factor $1/\sqrt{n}$ compared to σ_X .

The Central Limit Theorem describes the asymptotic fluctuations of \bar{X} around $\mathbb{E}(X)$:

Theorem 8.13. If X_1, X_2, \dots are independent with common distribution $X_i \sim X$, and if $\mathbb{E}(X^2) < \infty$ and $\text{Var}(X) \neq 0$, then the distribution of $\frac{\sqrt{n}}{\sqrt{\text{Var}(X)}}(\bar{X} - \mathbb{E}(X))$ converges as $n \rightarrow \infty$ to the distribution of a standard Normal random variable, in the sense that

$$\lim_{n \rightarrow \infty} \mathbb{P} \left(\frac{\bar{X} - \mathbb{E}(X)}{\sqrt{\text{Var}(X)/n}} \leq z_0 \right) = \int_{-\infty}^{z_0} \frac{e^{-z^2/2}}{\sqrt{2\pi}} dz$$

for any $z_0 \in \mathbb{R}$.

These two theorems say that the sample mean \bar{X} concentrates around the true expected value $\mathbb{E}(X)$ when n is large, and indeed the distribution of \bar{X} resembles a narrow bell curve centred around $\mathbb{E}(X)$. Consequently, if the assumptions about X_1, \dots, X_n and X are valid, then the sample mean \bar{X} is a good estimator of the true expectation $\mathbb{E}(X)$.

If we want to estimate other properties of the unknown distribution X , the situation is more complicated. For instance, to estimate $\text{Var}(X)$, we might try the conditional variance of \hat{X} given the data:

$$\text{Var} \left(\hat{X} \mid X_1, \dots, X_n \right) = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}.$$

However, this quantity is a *biased estimator* of $\text{Var}(X)$ in the sense that

$$\mathbb{E} \left(\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n} \right) \neq \text{Var}(X).$$

In fact, Theorem 8.11 can be used to show that

$$\mathbb{E} \left(\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n} \right) = \frac{n-1}{n} \text{Var}(X)$$

and that an unbiased estimator of $\text{Var}(X)$ is given by the *sample variance*

$$\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}.$$

An important part of the field of Statistics is concerned with finding good estimators for parameters in more complicated models.

8.6.4 Other relationships between data and randomness

Bootstrapping According to Theorem 8.12, the Strong Law of Large Numbers, the sample mean \bar{X} is a better estimator of the true mean $\mathbb{E}(X)$ if n , the number of data values, is large. In short, other things being equal, having more data values is better. Suppose we have an existing data set X_1, X_2, \dots, X_n . Can we simulate “new” data values – but without actually making any new observations – to get a better estimator for the mean?

The short answer is no: you cannot get something for nothing. However, although we cannot improve the estimator \bar{X} , we can use new simulated values to make guesses about the sampling distribution of \bar{X} . This is the basic idea of *bootstrapping*.

When we use the sample mean as an estimator for the true expectation, we are substituting the experiment “choose, uniformly at random, a data value from a given data set” for the unknown experiment “observe a new, previously unobserved, data value.” That is, we are using the conditional distribution of \hat{X} (given the data X_1, X_2, \dots, X_n) as a proxy for the unknown distribution of X . Theorem 8.12, the Strong Law of Large Numbers, asserts that this is reasonable, at least at the level of expected values, if n is large.

The idea of bootstrapping is to repeat this substitution many times. Let N_1, N_2, \dots, N_n be mutually independent and independent of X_1, \dots, X_n , with $N_i \sim \text{DU}(n)$ for all i . Define

$$\hat{X}_i = X_{N_i}.$$

That is, each value \hat{X}_i is chosen uniformly at random from the data values X_1, \dots, X_n , and the choice of which one – the random index N_i – is independent for different i . However, the same data values X_1, \dots, X_n are used for all values of i . This gives us n “bootstrapped” data values $\hat{X}_1, \hat{X}_2, \dots, \hat{X}_n$. As mentioned, we expect that each \hat{X}_i has a conditional distribution (given the data values X_1, \dots, X_n) that is a fair approximation of the unknown distribution of X . So if we define

$$\hat{\bar{X}} = \frac{\hat{X}_1 + \hat{X}_2 + \dots + \hat{X}_n}{n},$$

then we can expect the distribution of $\hat{\bar{X}}$ to be close to the distribution of \bar{X} . Using a computer, we can repeat this procedure and simulate $\hat{\bar{X}}$ as many times as we want, at low cost. That is, bootstrapping gives a convenient way to estimate the *sampling distribution* of \bar{X} .

Note that bootstrapping gives an approximation to the true distribution of X , and to the sampling distribution of \bar{X} . However, this is only an approximation and is subject to limitations. For instance, since bootstrapping only recycles the n data values obtained in the original sample, there is no way to find out about extreme values and outliers.

Bayesian statistics Our model above assumes that the data values X_1, X_2, \dots, X_n are independent copies of a random variable X . This is a *frequentist* model: parameters of interest, such as $\mathbb{E}(X)$, are assumed to be fixed even though they are unknown, and we look for estimators that will give values close to $\mathbb{E}(X)$ regardless of what the value of $\mathbb{E}(X)$ actually is.

In *Bayesian statistics*, the uncertainty in the parameters is incorporated into the probability model. We start with a parameter Θ that is a random variable; the distribution of Θ is called the *prior distribution*. The prior distribution encodes our existing beliefs, if any, about plausible values of the parameter. Often, we have no existing beliefs about parameter values, and we choose a “wide” prior distribution that assigns at least a moderate amount of probability to all possible parameter values.

The data are modelled by specifying the conditional distribution of X_1, X_2, \dots, X_n given Θ . Given the data, we then estimate Θ using the *posterior distribution*, the conditional distribution of Θ given X_1, \dots, X_n . Bayesian statistics is therefore a systematic way of updating our beliefs about Θ in response to observed data X_1, \dots, X_n . Typically, our updated beliefs becomes much more specific when we obtain data: the posterior distribution for Θ given X_1, X_2, \dots, X_n becomes *concentrated* in a narrow window around just one possible value of Θ .

8.7 Exercises

8.7.1 Show that the definition $\text{Cov}(X, Y) = \mathbb{E} \left[(X - \mathbb{E}[X]) (Y - \mathbb{E}[Y]) \right]$ is the same as the shortcut formula $\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$.

8.7.2 Find $\text{Cov}(X, Y^X)$ and ρ_{X, Y^X} , where X and Y are independent random variables with $\mathbb{P}(X = 1) = \mathbb{P}(X = 2) = \mathbb{P}(Y = 1) = \mathbb{P}(Y = 2) = 1/2$.

8.7.3 Let X and Y be random variables (neither one of which is constant) and define

$$\hat{Y} = \mathbb{E}(Y) + \frac{\text{Cov}(X, Y)}{\text{Var}(X)}(X - \mathbb{E}(X)).$$

1. Show that $\mathbb{E}(\hat{Y}) = \mathbb{E}(Y)$ and $\text{Cov}(X, \hat{Y}) = \text{Cov}(X, Y)$.

2. Show that

$$\frac{\text{Var}(\hat{Y})}{\text{Var}(Y)} = \rho_{X, Y}^2.$$

8.7.4 Let X and Y be random variables. Among all predictions of the form $\hat{Y} = aX$ (with no constant term), find the value of a that minimises the mean squared error of prediction $\mathbb{E}((Y - \hat{Y})^2)$.

8.7.5 Suppose that random variables X and Y satisfy $\mathbb{E}(Y | X) = 0$ (i.e., $\mathbb{E}(Y | X = x) = 0$ for any x for which $f_X(x) > 0$). Show that $\text{Cov}(X, Y) = 0$.

9 Introduction to Markov chains

By the end of this chapter you should be able to:

- formulate the Markov property for a randomly evolving system
- convert between a transition diagram and a transition matrix
- use the transition matrix to compute n -step transition probabilities

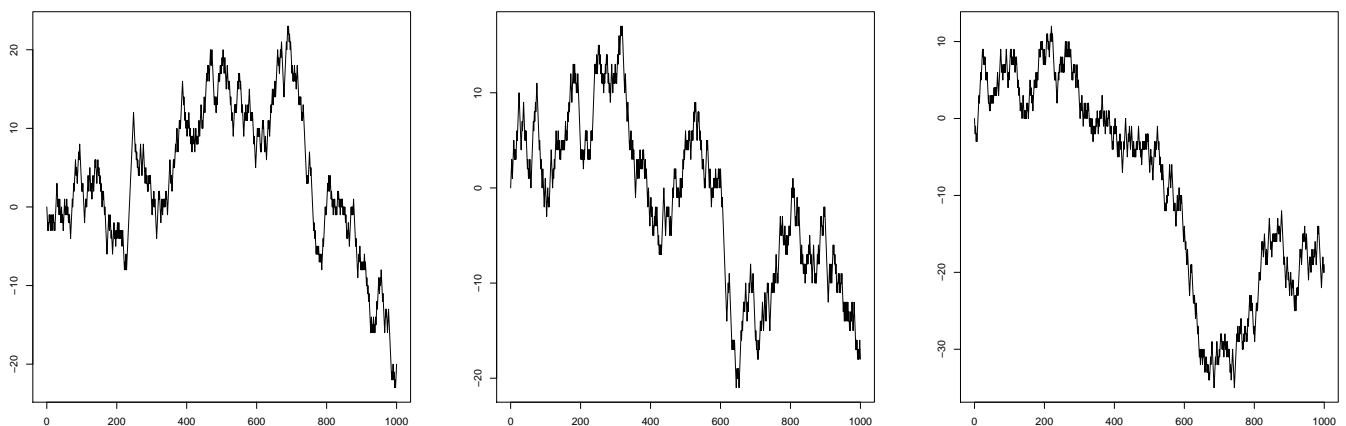
9.1 Examples

Example: Random Walks

(a) Simple Random Walk: at times $1, 2, 3, \dots$, throw a coin

If the coin lands on Heads, jump *up* 1 step (with probability 0.5).
If the coin lands on Tails, jump *down* 1 step (with probability 0.5).

Three realizations of 1000 steps of the simple random walk.
In all of the following, the x -axis is time and the y -axis is the state.



(b) Asymmetric random walk: Use an unfair coin.

Jump up with probability (w.p.) p
Jump down w.p. $1 - p$.

(c) Gambler's ruin

Start with, for instance, \$50.
Bet \$5 on the throw of a coin.
If it comes up heads win \$5, otherwise lose \$5.
Heads occur with probability p .
Stop when reach \$0 (bankruptcy) or \$100.

This is known as the random walk with absorbing barriers.

Other examples:
reflecting barriers
varying jump sizes

Example: Size of a threatened bird population

We might wish to model the number of kiwis in a bird sanctuary.

Deaths = decrease.
Births = increase.

We record the number of kiwis present in the sanctuary every week.

Probability death occurs $\approx ap$ where p is small and a is size of population.

Probability birth occurs? Depends on time of year.

In general, our simple models will not allow probabilities of death, birth, H, T, etc. to depend on how long the process has been running.

Example: Size of a queue

Individuals join a queue at rate λ per unit time and are served at rate μ per unit time.

Other examples of processes that evolve in a stochastic (random) fashion over time include:

Stock market
:
Bus arrivals
Lotto - Jackpot
Traffic
Internet
Supermarket queues
Weather

A *stochastic process* is a process that evolves randomly over time. More formally, a stochastic process is a collection of random variables indexed by time.

Time for us will be discrete e.g. throws of a die, hours, days, weeks (but it could also be continuous).

Definition: A discrete-time stochastic process is a sequence (X_0, X_1, X_2, \dots) indexed by time $n = 0, 1, 2, \dots$. The random variable X_n is called the *state* of the system at time n . The set of possible states of a stochastic process is called the *state space* and denoted S .

[Previously, we used S to denote the sample space for a random experiment. This is not the same as the state space. The states are the possible values of the process X_n at one specific time, whereas a sample point must include enough information to encode the entire random experiment, including the values X_n at every possible time. Henceforth, we will use S to denote the state space, and we will avoid mentioning the sample space.]

In this course, the state space of the stochastic process will also be discrete, i.e., each X_n will be a discrete random variable.

Example: In a random walk, the possible states are $\dots, -2, -1, 0, 1, 2, \dots$. The state space is $S = \mathbb{Z}$, the set of integers. The stochastic process changes by ± 1 at every step.

Example: In a bird population, the possible states are $0, 1, 2, \dots$. The state space is $S = \mathbb{N} \cup \{0\}$, the set of non-negative integers. At every step, the stochastic process can remain the same, or change to any other state, depending on the number of births and deaths.

We need to make some assumptions about how the system behaves.

The simplest assumption is to assume that

$$X_0, X_1, \dots$$

are all independent of one another, that is,

$$\mathbb{P}(X_{n+1} = s_{n+1} | X_n = s_n, \dots, X_0 = s_0) = \mathbb{P}(X_{n+1} = s_{n+1})$$

but this doesn't give us a very interesting system!

9.2 The Markov property

Definition: A stochastic process (X_0, X_1, X_2, \dots) satisfies the *Markov property* if

$$\mathbb{P}(X_{n+1} = s_{n+1} | X_n = s_n, \dots, X_0 = s_0) = \mathbb{P}(X_{n+1} = s_{n+1} | X_n = s_n)$$

for any states s_0, \dots, s_n, s_{n+1} .

Interpretation in words:

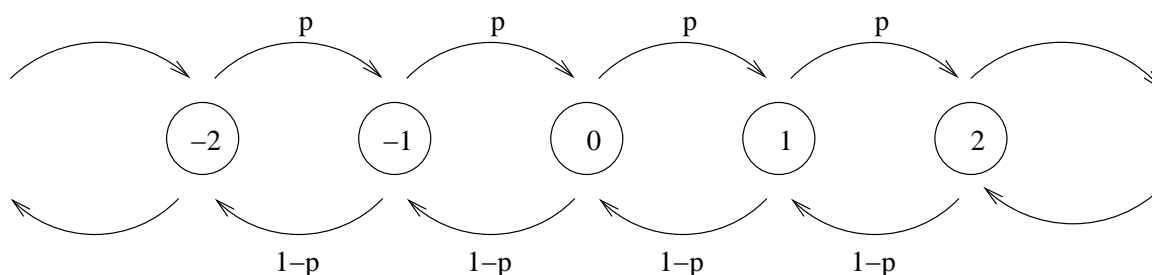
What happens next depends only on the current state of the system and not on its prior history.

This is called the *Markov property* and a discrete-time process with this property is called a *Markov chain*.

Example: the random walk

$\mathbb{P}(X_{n+1} = a + 1 | X_n = a) = p$ (the prob. of getting a head on the $(n + 1)$ st toss of the coin)

$\mathbb{P}(X_{n+1} = a - 1 | X_n = a) = 1 - p$.



Example: a 2-state Markov chain with states labelled 1 and 2, so $S = \{1, 2\}$.

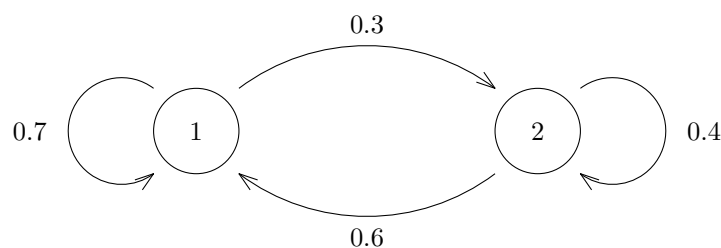
Suppose $\mathbb{P}(X_{n+1} = 2 | X_n = 1) = p_{12} = 0.3$

$\mathbb{P}(X_{n+1} = 1 | X_n = 1) = 1 - p_{12} = 0.7$

$\mathbb{P}(X_{n+1} = 1 | X_n = 2) = p_{21} = 0.6$

$\mathbb{P}(X_{n+1} = 2 | X_n = 2) = 1 - p_{21} = 0.4$.

We can draw a *transition diagram* that summarizes this information.



Here the size of the state space is $|S| = 2$.

We can also summarize this information in a *matrix*.

9.3 Transition matrices

A matrix is an array of entries organised in rows and columns. The entry in the i^{th} row and j^{th} column has subscripts ij .

For our 2-state Markov chain

$$P = \begin{pmatrix} 1 - p_{12} & p_{12} \\ p_{21} & 1 - p_{21} \end{pmatrix}$$

If we write $1 - p_{12} = p_{11}$, $1 - p_{21} = p_{22}$ then

$$P = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix}$$

Note that $p_{11} + p_{12} = 1$ and $p_{21} + p_{22} = 1$.

$$P = \begin{pmatrix} 0.7 & 0.3 \\ 0.6 & 0.4 \end{pmatrix}$$

More generally:

Definition:

$$p_{ij} = \mathbb{P}(X_{n+1} = j | X_n = i)$$

is the *transition probability from state i to state j* .

Our notation (p_{ij} with no n) assumes that the transition probability $\mathbb{P}(X_{n+1} = j | X_n = i)$ does not depend on n . We say the chain is *time-homogeneous* if that is the case. We will only consider time-homogeneous chains.

Definition: The *transition matrix* associated to a Markov chain with state space $S = \{1, 2, \dots, n\}$ and transition probabilities ($p_{ij}, i, j = 1, \dots, n$) is the $n \times n$ matrix P with p_{ij} in the i^{th} row and j^{th} column.

We always have $p_{ij} \geq 0$ for all i, j in S and from each state i the transition probabilities add to 1, i.e.

$$\sum_{j \in S} p_{ij} = 1.$$

This means that

- The entries of P are non-negative; and
- The sum along any *row* of P equals 1.

Note: In usual matrix notation, the rows and columns of a matrix are counted starting from 1. When $S = \{1, 2, \dots, n\}$, this means that the i^{th} row contains transition probabilities starting from state i , and the j^{th} column contains transition probabilities ending in state j .

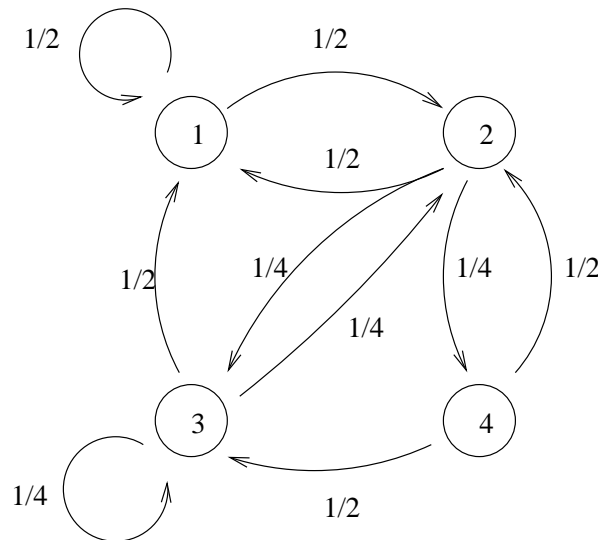
Often the state space is $S = \{0, 1, 2, \dots, k\}$, with the state 0 included. In this case, it makes sense to count the rows and columns starting from 0 instead, even though this conflicts usual matrix notation. We will see examples later. When this happens, remember that the topmost row, and leftmost column, correspond to state 0 instead of state 1, and the matrix has $k + 1$ rows and $k + 1$ columns.

In general, we can always write down the transition probabilities in a matrix, provided that we order the rows and columns in the same way. The transition matrix P will have size $|S| \times |S|$ in all cases.

Example:

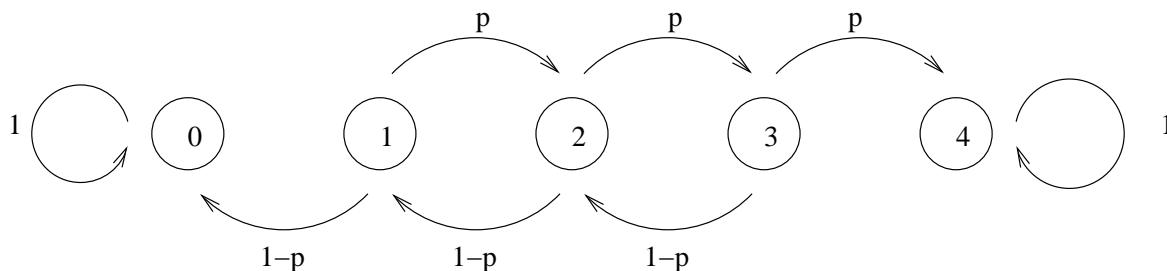
Draw the transition diagram for the Markov chain with transition matrix

$$P = \begin{pmatrix} 0.5 & 0.5 & 0 & 0 \\ 0.5 & 0 & 0.25 & 0.25 \\ 0.5 & 0.25 & 0.25 & 0 \\ 0 & 0.5 & 0.5 & 0 \end{pmatrix}.$$



Example:

Write down the transition matrix for the Markov chain with the following transition diagram.



$$P = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1-p & 0 & p & 0 & 0 \\ 0 & 1-p & 0 & p & 0 \\ 0 & 0 & 1-p & 0 & p \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Summary

Definition: The process $\{X_n; n = 0, 1, 2, \dots\}$ is a Markov chain if it satisfies the Markov property:

$$\mathbb{P}(X_{n+1} = s_{n+1} | X_0 = s_0, \dots, X_n = s_n) = \mathbb{P}(X_{n+1} = s_{n+1} | X_n = s_n)$$

for all $n = 0, 1, 2, 3, \dots$ and all s_0, s_1, \dots, s_{n+1} in S .

Definition: The transition probabilities of a Markov chain are

$$p_{ij} = \mathbb{P}(X_{n+1} = j | X_n = i)$$

for i, j in S , $n = 0, 1, 2, \dots$

Definition: The transition matrix $P = (P_{ij})$ is the $|S| \times |S|$ matrix of transition probabilities from state to state. The rows are the “from” states, and the columns are the “to” states (arranged in a sensible and consistent order).

9.4 Sample path behaviour and n -step transition probabilities

We would like to be able to say more about the evolution of a Markov chain — its *sample path* behaviour. For instance, what is

$$\begin{aligned} &\mathbb{P}(X_1 = 1, X_2 = 2, X_3 = 1 \mid X_0 = 0) \text{ or} \\ &\mathbb{P}(X_2 = 3 \mid X_0 = 1) \text{ or} \\ &\mathbb{P}(X_{10} = 2 \mid X_3 = 7)? \end{aligned}$$

Let us begin by thinking about a single sample path.

Example: For the asymmetric random walk, find

$$\mathbb{P}(X_1 = 1, X_2 = 2, X_3 = 1, X_4 = 0, X_5 = -1, X_6 = -2 \mid X_0 = 0).$$

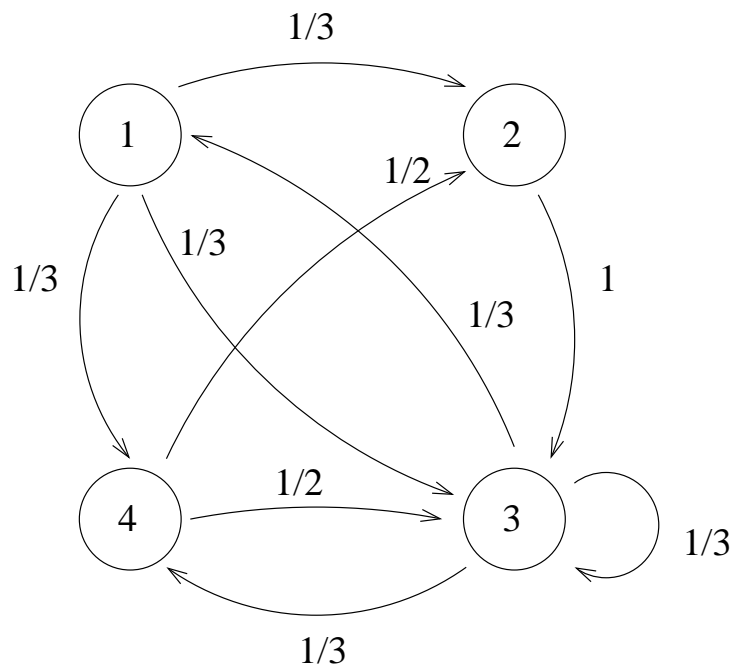
Using conditioning and the Markov property,

$$\begin{aligned} \mathbb{P}(X_1 = 1, X_2 = 2 \mid X_0 = 0) &= \mathbb{P}(X_1 = 1 \mid X_0 = 0) \mathbb{P}(X_2 = 2 \mid X_1 = 1, X_0 = 0) \\ &= \mathbb{P}(X_1 = 1 \mid X_0 = 0) \mathbb{P}(X_2 = 2 \mid X_1 = 1) = p_{01}p_{12}. \end{aligned}$$

This pattern repeats for subsequent steps. Since the probability of taking an up step is $p_{i,i+1} = p$, the answer is

$$p \cdot p \cdot (1 - p) \cdot (1 - p) \cdot (1 - p) \cdot (1 - p) = p^2(1 - p)^4.$$

Example:



What is the transition matrix

$$P = \begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & 1 & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}$$

and what is the sample path probability

$$\mathbb{P}(X_1 = 4, X_2 = 2, X_3 = 3, X_4 = 1 \mid X_0 = 1)?$$

$$\begin{aligned} \mathbb{P}(X_1 = 4, X_2 = 2, X_3 = 3, X_4 = 1 \mid X_0 = 1) &= p_{14}p_{42}p_{23}p_{31} \\ &= \frac{1}{3} \cdot \frac{1}{2} \cdot 1 \cdot \frac{1}{3} \\ &= \frac{1}{18} \end{aligned}$$

In general, we have

$$\mathbb{P}(X_1 = s_1, X_2 = s_2, \dots, X_n = s_n \mid X_0 = s_0) = p_{s_0 s_1} p_{s_1 s_2} \cdots p_{s_{n-1} s_n}$$

for any $n = 1, 2, \dots$ and $s_0, s_1, s_2, \dots, s_n$ in S .

That is, the probability of observing a particular sample path can be found by multiplying together the appropriate transition probabilities from the transition matrix.

This tells us how to find the probability of a particular sample path.

What if we wish to find something like

$$\mathbb{P}(X_2 = 3 \mid X_0 = 1)?$$

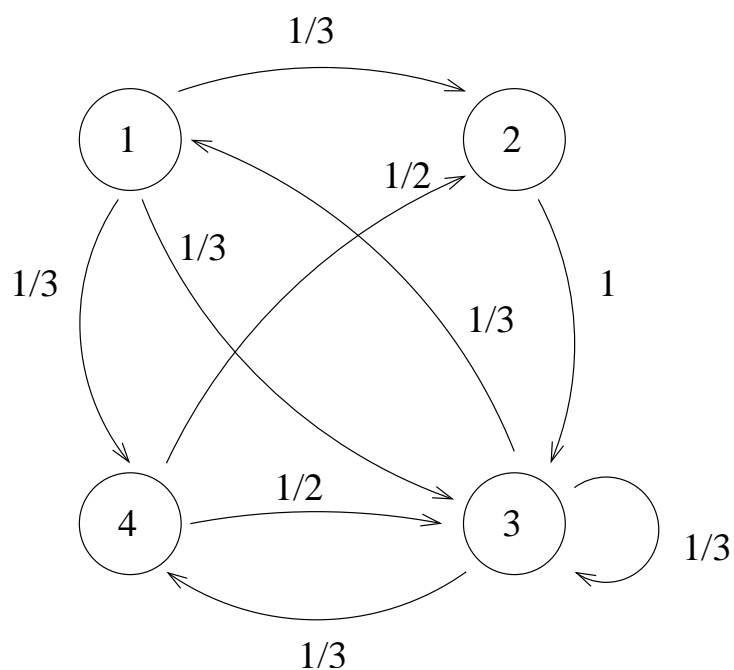
We will write this as $p_{13}^{(2)}$ and this is an example of an n -step transition probability with $n = 2$.

In general, we call

$$\mathbb{P}(X_n = j \mid X_0 = i) = p_{ij}^{(n)}$$

the n -step transition probability from i to j .

Example:



What is $p_{13}^{(2)}$?

Possible ways of getting from 1 to 3 in 2 steps:

$$\left. \begin{array}{lll} 1 \rightarrow 2 \rightarrow 3 & \text{w. p.} & \frac{1}{3} \cdot 1 = \frac{1}{3} \\ 1 \rightarrow 3 \rightarrow 3 & \text{w. p.} & \frac{1}{3} \cdot \frac{1}{3} = \frac{1}{9} \\ 1 \rightarrow 4 \rightarrow 3 & \text{w. p.} & \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6} \end{array} \right\} \begin{array}{l} \frac{1}{3} + \frac{1}{9} + \frac{1}{6} \\ = \frac{6+2+3}{18} = \frac{11}{18} \end{array}$$

What about $p_{12}^{(2)}$?

Only possible path is $1 \rightarrow 4 \rightarrow 2$ w. p. $\frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6}$.

Note: "w.p." is the abbreviation of "with probability".

Find $p_{11}^{(2)}$ and $p_{14}^{(2)}$, and check that $p_{11}^{(2)} + p_{12}^{(2)} + p_{13}^{(2)} + p_{14}^{(2)} = 1$.

Now suppose that we wish to find

$$p_{13}^{(3)} = \mathbb{P}(X_3 = 3 \mid X_0 = 1).$$

How can we get from 1 to 3 in 3 steps?

$1 \rightarrow 2 \rightarrow 3 \rightarrow 3$
 $1 \rightarrow 3 \rightarrow 3 \rightarrow 3$
 $1 \rightarrow 4 \rightarrow 3 \rightarrow 3$
 $1 \rightarrow 3 \rightarrow 1 \rightarrow 3$
 $1 \rightarrow 4 \rightarrow 2 \rightarrow 3$
 $1 \rightarrow 3 \rightarrow 4 \rightarrow 3$

More precisely, what we are doing is writing

$$\begin{aligned}
 p_{13}^{(3)} &= \mathbb{P}(X_3 = 3 \mid X_0 = 1) \\
 &= \mathbb{P}(X_3 = 3, X_2 = 3, X_1 = 2 \mid X_0 = 1) \\
 &\quad + \mathbb{P}(X_3 = 3, X_2 = 3, X_1 = 3 \mid X_0 = 1) \\
 &\quad + \mathbb{P}(X_3 = 3, X_2 = 3, X_1 = 4 \mid X_0 = 1) \\
 &\quad + \mathbb{P}(X_3 = 3, X_2 = 1, X_1 = 3 \mid X_0 = 1) \\
 &\quad + \mathbb{P}(X_3 = 3, X_2 = 2, X_1 = 4 \mid X_0 = 1) \\
 &\quad + \mathbb{P}(X_3 = 3, X_2 = 4, X_1 = 3 \mid X_0 = 1).
 \end{aligned}$$

We already know that each of these quantities is a product of transition probabilities, so what we are doing here is adding up products of transition probabilities. This is tedious work - the kind of thing that computers are useful for!

Such expressions are often written using matrix notation. Consider

$$P \times P = \begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & 1 & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix} \times \begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & 1 & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}$$

If we compute the matrix $P^2 = P \times P$, the entry $(P^2)_{i,j}$ is the probability of going from state i to state j in 2 steps. For our example, $p_{13}^{(2)}$ is the entry in row 1, column 3 of the matrix P^2 .

Definition: For two square $n \times n$ matrices, A and B , then the ij^{th} entry in $C = A \times B$ is found by

$$c_{ij} = \sum_{k=1}^n a_{ik} b_{kj}.$$

So in this example, since P is arranged so that $P_{ij} = p_{ij}$, $p_{13}^{(2)}$ is obtained by taking the 1, 3 entry in the matrix P^2 , that is, $p_{13}^{(2)} = (P^2)_{13}$.

In general, if P is arranged so that $P_{ij} = p_{ij}$, then $p_{ij}^{(2)}$ is obtained by taking the ij^{th} entry in P^2 . So $P^{(2)} = P^2$.

Similarly $P^{(3)} = P^{(2)} \times P = P \times P \times P = P^3$.

In general, $P^{(n)} = \underbrace{P \times \dots \times P}_{n \text{ times}} = P^n$.

For our example

$$P^{(2)} = \begin{pmatrix} \frac{1}{9} & \frac{1}{6} & \frac{11}{18} & \frac{1}{9} \\ \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{9} & \frac{5}{18} & \frac{7}{18} & \frac{2}{9} \\ \frac{1}{6} & 0 & \frac{2}{3} & \frac{1}{6} \end{pmatrix}$$

and

$$P^{(3)} = \begin{pmatrix} \frac{1}{9} & \frac{1}{6} & \frac{11}{18} & \frac{1}{9} \\ \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{9} & \frac{5}{18} & \frac{7}{18} & \frac{2}{9} \\ \frac{1}{6} & 0 & \frac{2}{3} & \frac{1}{6} \end{pmatrix} \times \begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & 1 & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}$$

Computers do these kinds of calculations for us!!

The most general statement of the preceding results is given by

The Chapman-Kolmogorov equations

$$p_{ij}^{(m+n)} = \sum_{k \in S} p_{ik}^{(m)} p_{kj}^{(n)}$$

for i, j in S , for $m, n = 1, 2, 3, \dots$

In matrix notation this is just

$$P^{(m+n)} = P^{(m)} \times P^{(n)}$$

In some cases, finding the n -step transition probabilities is relatively straightforward.

Example: The random walk again.

What is $p_{00}^{(n)}$?

For the random walk, the steps taken (“up” vs. “down”) form a sequence of Bernoulli trials. (Reason: An up-step always has the same conditional probability, regardless of the current state [because the transition probabilities $p_{i,i+1}$ are the same for every i] or the previous history [because of the Markov property].) Thus the number of up-steps at time n has the Binomial(n, p) distribution.

$p_{00}^{(n)} = \mathbb{P}(X_n = 0 | X_0 = 0)$ means the probability that the number of up-steps equals the number of down-steps. The total number of steps is n , so this only works if n is even and

$$p_{00}^{(n)} = \mathbb{P}(\#(\text{up-steps}) = \frac{n}{2}) = \mathbb{P}(\text{Binomial}(n, p) = \frac{n}{2}) = \begin{cases} \binom{n}{n/2} p^{n/2} (1-p)^{n-n/2} & \text{if } n \text{ is even,} \\ 0 & \text{if } n \text{ is odd.} \end{cases}$$

Summary

For any Markov chain,

$$\mathbb{P}(X_1 = s_1, X_2 = s_2, \dots, X_n = s_n | X_0 = s_0) = p_{s_0 s_1} p_{s_1 s_2} \dots p_{s_{n-1} s_n}$$

for any $n = 1, 2, \dots$ and $s_0, s_1, s_2, \dots, s_n$ in S .

Definition

The n -step probability from i to j is $\mathbb{P}(X_n = j | X_0 = i) = p_{ij}^{(n)}$.

$$P^{(n)} = \underbrace{P \times \dots \times P}_{n \text{ times}} = P^n.$$

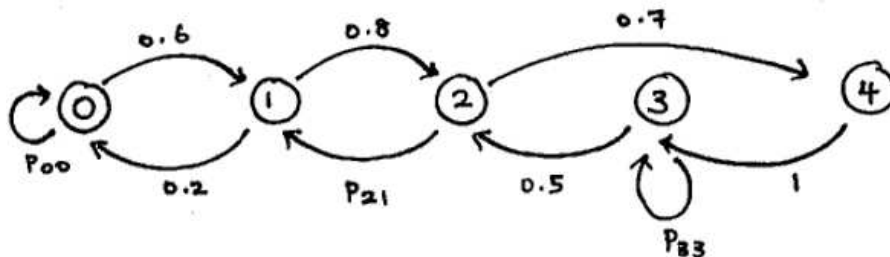
The Chapman-Kolmogorov equations

$$p_{ij}^{(m+n)} = \sum_{k \in S} p_{ik}^{(m)} p_{kj}^{(n)}$$

for i, j in S , for $m, n = 1, 2, 3, \dots$

9.5 Exercises

9.5.1 A Markov chain has the following transition diagram



1. What is the state space for this Markov chain?
2. What are the values of p_{00} , p_{21} and p_{33} ?
3. Give the sensible transition matrix P for this Markov chain.

9.5.2 A Markov chain $\{X_n, n = 0, 1, 2, \dots\}$ with states $\{0, 1, 2\}$ has transition matrix

$$\begin{pmatrix} 0.1 & 0.2 & 0.7 \\ 0.9 & 0.1 & 0.0 \\ 0.1 & 0.8 & 0.1 \end{pmatrix}$$

and initial distribution $\mathbb{P}(X_0 = 0) = 0.3$, $\mathbb{P}(X_0 = 1) = 0.4$ and $\mathbb{P}(X_0 = 2) = 0.3$. Find the following

1. $\mathbb{P}(X_0 = 0, X_1 = 2, X_2 = 2)$
2. $p_{0i}^{(2)} = \mathbb{P}(X_2 = i | X_0 = 0)$ for $i = 0, 1, 2$.
3. $\mathbb{P}(X_3 = 1 | X_1 = 0)$
4. $\mathbb{P}(X_3 = 1 | X_0 = 0)$.

9.5.3 Consider the asymmetric random walk with transition probabilities $p_{i,i+1} = p$,
 $p_{i,i-1} = 1 - p = q$ for $i \in \mathbb{Z}$, and $p_{ij} = 0$ otherwise.

Find $\mathbb{P}(X_n \geq 0, n = 1, 2, 3 | X_0 = 0) = \mathbb{P}(X_1 \geq 0, X_2 \geq 0, X_3 \geq 0 | X_0 = 0)$.

10 Equilibrium and limiting distributions

By the end of this chapter you should be able to:

- verify whether a distribution is an equilibrium distribution
- for a given Markov chain, determine whether an equilibrium distribution exists and compute it using the Full Balance Equations
- use the Detailed Balance Equations, where appropriate, as an aid to finding the equilibrium distribution
- apply general criteria for the existence and uniqueness of equilibrium distributions and limiting distributions

10.1 Equilibrium distributions and the Full Balance Equations

In a Markov chain, there are two ingredients:

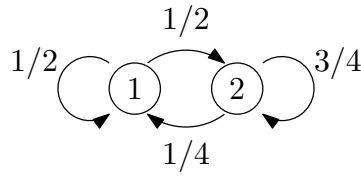
- the *transition probabilities* – the conditional probabilities p_{ij} , $i, j \in S$ of moving to the various states j at the next step, given the current state i ; and
- the *initial distribution* – the probabilities $\mathbb{P}(X_0 = i)$, $i \in S$, of being at particular states initially.

We can specify these two ingredients separately, depending on what situation we want to model or study. We might specify a fixed starting state $i \in S$: this corresponds to setting $\mathbb{P}(X_0 = i) = 1$ and $\mathbb{P}(X_0 = i') = 0$ for $i' \neq i$. Or we might assign X_0 randomly: for instance, we might declare that the initial state X_0 should be uniformly distributed across all possible states (provided the state space S is finite). In either case we are interested in the distribution of X_n at later times n – i.e., the probabilities $\mathbb{P}(X_n = j)$ for $j \in S$. If we chose a fixed starting point i , these probabilities are precisely the n -step transition probabilities $p_{ij}^{(n)}$, $j \in S$, and generally the Partition Theorem gives

$$\begin{aligned}\mathbb{P}(X_1 = j) &= \sum_{i \in S} \mathbb{P}(X_0 = i) p_{ij}, \\ \mathbb{P}(X_{n+1} = j) &= \sum_{i \in S} \mathbb{P}(X_n = i) p_{ij} \quad \mathbb{P}(X_n = j) = \sum_{i \in S} \mathbb{P}(X_0 = i) p_{ij}^{(n)}\end{aligned}\tag{*}$$

Typically, the distribution of X_1 will not match the distribution of X_0 . For instance, even if we fix a starting point $i \in S$ (so that $\mathbb{P}(X_0 = i) = 1$ and the random variable X_0 is actually constant), the random variable X_1 will generally be non-constant due to the random jumping of the Markov chain. Furthermore, X_2 will typically have a *different* distribution from X_1 , as the Markov chain starts to reach states that were unreachable, or less likely to be reached, from the initial state i .

Example 10.1. Consider the Markov chain with the following transition diagram:



Suppose the Markov chain is started from state 1, i.e., our initial distribution is $\mathbb{P}(X_0 = 1) = 1, \mathbb{P}(X_0 = 2) = 0$. What are the distributions of X_1, X_2, X_3, X_4, X_5 ?

We need to find $\mathbb{P}(X_n = j)$ for each $j \in \{1, 2\}$ and each $n \in \{1, 2, 3, 4, 5\}$. Since there are only 2 states, $\mathbb{P}(X_n = 2) = 1 - \mathbb{P}(X_n = 1)$, so it is enough to find $\mathbb{P}(X_n = 1)$. Applying Equation (*) above,

$$\begin{aligned}
 \mathbb{P}(X_1 = 1) &= \mathbb{P}(X_0 = 1)p_{11} + \mathbb{P}(X_0 = 2)p_{21} \\
 &= (1)(1/2) + (0)(1/4) \\
 &= 1/2 = 0.5
 \end{aligned}$$

We can check that

$$\begin{aligned}
 \mathbb{P}(X_1 = 2) &= \mathbb{P}(X_0 = 1)p_{12} + \mathbb{P}(X_0 = 2)p_{22} \\
 &= (1)(1/2) + (0)(3/4) \\
 &= 1/2 = 1 - 1/2,
 \end{aligned}$$

as expected. Repeating this calculation,

$$\begin{aligned}
 \mathbb{P}(X_2 = 1) &= \mathbb{P}(X_1 = 1)(1/2) + \mathbb{P}(X_1 = 2)(1/4) \\
 &= (1/2)(1/2) + (1 - 1/2)(1/4) \\
 &= 3/8 = 0.375, \\
 \mathbb{P}(X_3 = 1) &= (3/8)(1/2) + (1 - 3/8)(1/4) \\
 &= 11/32 = 0.34375, \\
 \mathbb{P}(X_4 = 1) &= (11/32)(1/2) + (1 - 11/32)(1/4) \\
 &= 43/128 = 0.3359375, \\
 \mathbb{P}(X_5 = 1) &= (43/128)(1/2) + (1 - 43/128)(1/4) \\
 &= 171/512 = 0.333984375.
 \end{aligned}$$

The probability of being in state 1 appears to decrease over time, and seems to be approaching $0.333333 \dots = 1/3$. The behaviour of the Markov chain seems to be *equilibrating* as the effect of the initial distribution is gradually forgotten.

What would happen if the Markov chain had started from $\mathbb{P}(X_0 = 1) = 1/3, \mathbb{P}(X_0 = 2) = 2/3$ instead?

$$\begin{aligned}
 \mathbb{P}(X_1 = 1) &= (1/3)(1/2) + (2/3)(1/4) = 1/3, \\
 \mathbb{P}(X_1 = 2) &= (1/3)(1/2) + (2/3)(3/4) = 2/3.
 \end{aligned}$$

So the distribution of X_1 would be the same as the distribution of X_0 , and by the same argument X_n will have the same distribution for all $n \geq 0$.

In general, we wish to answer the following questions:

- What starting distributions for X_0 will lead to the same distribution for X_1 ?
- Will the distribution of X_n approach some fixed distribution when n is large (i.e., after running the Markov chain for a long time)?

These kinds of distributions will be called *equilibrium* and *limiting* distributions, respectively. We also want to know whether such distributions always exist, and whether there can be many of them.

Definition: An *equilibrium* (or *stationary* or *steady-state* or *invariant*) distribution is a vector of probabilities

$$(\pi_i, i \in S) \quad \text{with} \quad \pi_i \geq 0 \text{ for all } i \in S \quad \text{and} \quad \sum_{i \in S} \pi_i = 1$$

such that if $\mathbb{P}(X_0 = i) = \pi_i$ for all $i \in S$, then $\mathbb{P}(X_1 = i) = \pi_i$ for all $i \in S$.

If $\mathbb{P}(X_0 = i) = \pi_i$ for all $i \in S$, then $\mathbb{P}(X_n = i) = \pi_i$ for all $n \geq 0$ and all $i \in S$, and the Markov chain is said to be *in equilibrium*.

The condition for an equilibrium distribution is that $\mathbb{P}(X_0 = i)$ and $\mathbb{P}(X_1 = i)$ should both equal π_i , for every $i \in S$. Because of Equation (*) from page 183, we can write this condition as a collection of equations, called the *Full Balance Equations* or *Global Balance Equations*:

Theorem 10.1 ((The Full Balance Equations)). The non-negative numbers $(\pi_i, i \in S)$ is an equilibrium distribution if and only if

$$\pi_j = \sum_{i \in S} \pi_i p_{ij} \quad \text{for all } j \in S$$

with the additional condition that

$$\sum_{i \in S} \pi_i = 1.$$

Example 10.2 (Example 10.1 continued). For the Markov chain with transition diagram shown on page 184, find all possible equilibrium distributions.

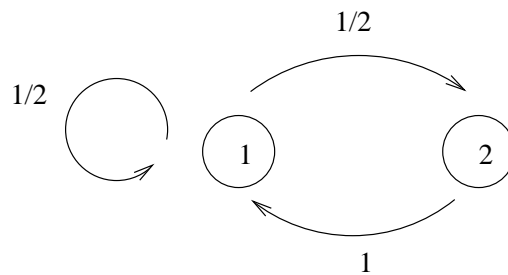
We set up the Full Balance Equations:

$$\begin{aligned} \pi_1 &= \pi_1 p_{11} + \pi_2 p_{21} = \pi_1(1/2) + \pi_2(1/4), \\ \pi_2 &= \pi_1 p_{12} + \pi_2 p_{22} = \pi_1(1/2) + \pi_2(3/4). \end{aligned}$$

Rearranging either equation simplifies to $(1/2)\pi_1 = (1/4)\pi_2$, so $\pi_2 = 2\pi_1$. To continue, use the additional condition $\pi_1 + \pi_2 = 1$ to get $3\pi_1 = 1$. Thus $\pi_1 = 1/3$ and $\pi_2 = 1 - 1/3 = 2/3$ gives the unique equilibrium distribution.

This matches our calculation before, where we showed that if $\mathbb{P}(X_0 = 1) = 1/3 = \pi_1$ and $\mathbb{P}(X_0 = 2) = 2/3 = \pi_2$ then X_1 had the same distribution.

Example 10.3. Consider the Markov chain with the following transition diagram:



Find all possible equilibrium distributions.

The Full Balance Equations are

$$\pi_1 = \pi_1(1/2) + \pi_2(1),$$

$$\pi_2 = \pi_1(1/2) + \pi_2(0),$$

$$1 = \pi_1 + \pi_2.$$

Either of the first two equations can be rearranged to say $\pi_2 = (1/2)\pi_1$, and therefore the third equation gives $1 = (3/2)\pi_1$. Hence $\pi_1 = 2/3$ and $\pi_2 = (1/2)(2/3) = 1/3$. Thus there is only one possible equilibrium distribution, in which state 1 has probability $2/3$ and state 2 has probability $1/3$.

Example 10.4. What is/are the equilibrium distribution(s) of the Markov chain with state space $S = \{0, 1, 2\}$ and transition matrix

$$P = \begin{pmatrix} 1/3 & 2/3 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 2/3 & 1/3 \end{pmatrix}?$$

The equilibrium distribution satisfies the equations

$$\begin{aligned}\pi_0 &= \frac{1}{3}\pi_0 + \frac{1}{2}\pi_1, \\ \pi_1 &= \frac{2}{3}\pi_0 + \frac{2}{3}\pi_2, \\ \pi_2 &= \frac{1}{2}\pi_1 + \frac{1}{3}\pi_2.\end{aligned}$$

with the additional condition that $\pi_0 + \pi_1 + \pi_2 = 1$.

From the first equation we obtain

$$\frac{2}{3}\pi_0 = \frac{1}{2}\pi_1 \Rightarrow \pi_0 = \frac{3}{4}\pi_1.$$

It is simpler to use the third equation (which only involves 2 of the three variables) instead of the second:

$$\frac{2}{3}\pi_2 = \frac{1}{2}\pi_1 \Rightarrow \pi_2 = \frac{3}{4}\pi_1.$$

With these two facts, the second equation becomes $\pi_1 = (2/4)\pi_1 + (2/4)\pi_1$, which is true but redundant.

Expressing the condition $\pi_0 + \pi_1 + \pi_2 = 1$ in terms of π_1 gives

$$1 = \left(\frac{3}{4} + 1 + \frac{3}{4}\right)\pi_1 = \frac{10}{4}\pi_1,$$

so

$$\pi_1 = \frac{4}{10}, \quad \pi_0 = \pi_2 = \frac{3}{10}.$$

These three numbers are uniquely determined, so there is only one equilibrium distribution.

Example 10.5. Find the equilibrium distribution(s) of the Markov chain with state space $S = \{0, 1, 2\}$ and transition matrix

$$P = \begin{pmatrix} 1-p & p & 0 \\ 1-p & 0 & p \\ 0 & 1-p & p \end{pmatrix}?$$

The equilibrium distribution satisfies the equations

$$\begin{aligned}\pi_0 &= (1-p)\pi_0 + (1-p)\pi_1 \\ \pi_1 &= p\pi_0 + (1-p)\pi_2 \\ \pi_2 &= p\pi_1 + p\pi_2\end{aligned}$$

with the additional condition that $\pi_0 + \pi_1 + \pi_2 = 1$.

From the first equation we obtain $p\pi_0 = (1-p)\pi_1$, i.e. $\Rightarrow \pi_1 = \frac{p}{1-p}\pi_0$. The third gives

$$(1-p)\pi_2 = p\pi_1 \Rightarrow \pi_2 = \frac{p}{1-p}\pi_1 \Rightarrow \pi_2 = \left(\frac{p}{1-p}\right)^2 \pi_0$$

So we have

$$\pi_1 = \frac{p}{1-p}\pi_0, \quad \pi_2 = \left(\frac{p}{1-p}\right)^2 \pi_0$$

and the condition $\pi_0 + \pi_1 + \pi_2 = 1$ then gives

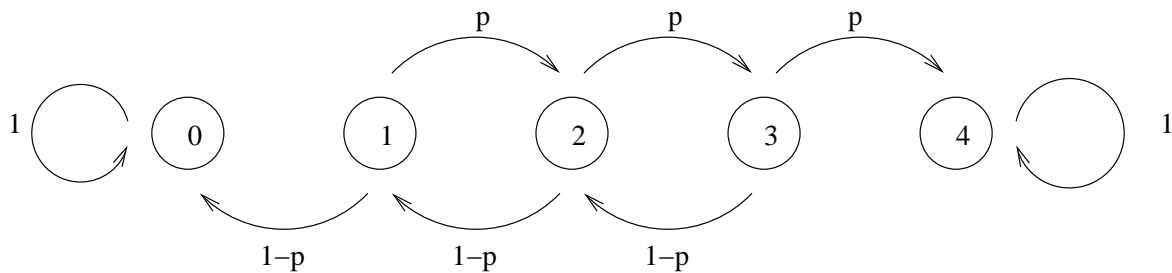
$$\pi_0 \left(1 + \frac{p}{1-p} + \left(\frac{p}{1-p}\right)^2 \right) = 1, \quad \text{i.e. } \pi_0 = \frac{1}{1 + \frac{p}{1-p} + \left(\frac{p}{1-p}\right)^2}.$$

From which we get

$$\pi_1 = \frac{\frac{p}{1-p}}{1 + \frac{p}{1-p} + \left(\frac{p}{1-p}\right)^2}, \quad \pi_2 = \frac{\left(\frac{p}{1-p}\right)^2}{1 + \frac{p}{1-p} + \left(\frac{p}{1-p}\right)^2}.$$

In all of the examples so far, there has been exactly one equilibrium distribution. However, this need not be the case.

Example 10.6. Find the equilibrium distribution(s) for the gambling Markov chain with transition diagram



The Full Balance Equations are

$$\pi_0 = \pi_0 + (1 - p)\pi_1$$

$$\pi_1 = (1 - p)\pi_2$$

$$\pi_2 = p\pi_1 + (1 - p)\pi_3$$

$$\pi_3 = p\pi_2$$

$$\pi_4 = p\pi_3 + \pi_4.$$

and $\pi_0 + \pi_1 + \pi_2 + \pi_3 + \pi_4 = 1$. The first equation gives $\pi_1 = 0$, whereupon the second and third equations give $\pi_2 = \pi_3 = 0$ also.

[This assumes $1 - p \neq 0$; if $p = 1$, the same conclusion is true, from looking at the fifth, fourth and third equations instead.]

However, once we set $\pi_1 = \pi_2 = \pi_3 = 0$, all five equations are satisfied for any values of π_0 and π_4 . The additional condition $\pi_0 + \pi_4 = 1$ gives $\pi_4 = 1 - \pi_0$ but does not uniquely specify π_0 and π_4 . In summary, as long as $0 \leq \pi_0 \leq 1$, setting $\pi_4 = 1 - \pi_0$ and $\pi_1 = \pi_2 = \pi_3 = 0$ gives an equilibrium distribution, and there are infinitely many possible equilibrium distributions.

In this example, all the probability mass is shared out between states 0 and 4. Starting from either of these states, the Markov chain will never move – the states are **absorbing** states – so any such distribution will be an equilibrium distribution.

Example 10.7. Let $p > 0$. Find the equilibrium distribution(s) for the Markov chain with state space $S = \{0, 1, 2, \dots\}$ and transitions described by

$$X_{n+1} = \begin{cases} X_n + 1 & \text{with probability } p, \\ X_n & \text{with probability } 1 - p. \end{cases}$$

The transition probabilities are $p_{i,i+1} = p$, $p_{ii} = 1 - p$. For each j , the states i contributing to the j^{th} Full Balance Equation are $i = j - 1$ and $i = j$ (since these are the only states i that can move to j in one step). So the Full Balance Equations are

$$\pi_j = \pi_{j-1}p_{j-1,j} + \pi_j p_{jj} = p\pi_{j-1} + (1-p)\pi_j$$

for all j . Rearranging, $p\pi_j = p\pi_{j-1}$ for all j , and since p is non-zero it must be that $\pi_j = \pi_{j-1}$ for all j . That is, π_j does not depend on j , so π_j must be some constant number, say x . It is impossible that $x = 0$, since in that case $\sum_{i \in S} \pi_i = 0$, violating the condition $\sum_{i \in S} \pi_i = 1$. On the other hand, it is impossible that $x > 0$, since in that case the sum $\sum_{i \in S} x$ is infinite, which also violates $\sum_{i \in S} \pi_i = 1$. So there can be no equilibrium distribution.

The result of Example 10.7 can only happen for Markov chains with infinitely many states:

Theorem 10.2. A Markov chain with a finite state space always has at least one equilibrium distribution.

10.2 Detailed Balance Equations

For a special but important class of Markov chains, we can use an easier method to find equilibrium distribution(s).

Definition: A Markov chain is called a *birth-death chain* if the state space S is $\{0, 1, 2, \dots\}$, or any subset of $\{0, 1, 2, \dots\}$, and all possible transitions are from i to $i + 1$, from i to $i - 1$, or from i to i .

[The transitions from i to $i + 1$, $i - 1$ or i do not all have to be possible.]

Theorem 10.3. If a Markov chain is a birth-death chain, any equilibrium distribution $(\pi_i, i \in S)$ must satisfy the equations

$$\pi_i p_{ij} = \pi_j p_{ji} \quad \text{for all } i, j \in S$$

We call these equations (over all pairs of states i and j) the **Detailed Balance Equations**.

Theorem 10.4. For any Markov chain, if the non-negative numbers $(\pi_i, i \in S)$ satisfy the Detailed Balance Equations and also satisfy $\sum_{i \in S} \pi_i = 1$, then $(\pi_i, i \in S)$ is an equilibrium distribution.

Remarks:

- The combination of Theorems 10.3 and 10.4 means that for a birth-death chain, we can find all equilibrium distributions (and determine whether any exist, when the state space is infinite) by solving the Detailed Balance Equations. Solving the Detailed Balance Equations is simpler than solving the Full Balance Equations since each Detailed Balance Equation involves only two unknowns.
- For a birth-death chain, the only relevant Detailed Balance Equations are those for which i and j differ by exactly 1 (the equation for $i = j$ is always satisfied, and the equations for $|i - j| \geq 2$ reduce to $0 = 0$ since both transition probabilities p_{ij} and p_{ji} are 0 in that case).
- For a Markov chain that is not a birth-death chain, Theorem 10.4 says that if we solve the Detailed Balance Equations then we automatically solve the Full Balance Equations. However, we cannot use the Detailed Balance Equations as a way of finding equilibrium distributions in general, since there is no guarantee that all equilibrium distributions will satisfy the Detailed Balance Equations.
- If a transition from i to j is possible ($p_{ij} > 0$) but the reverse transition is not ($p_{ji} = 0$) then the Detailed Balance Equations require that $\pi_i = 0$. In particular, if this is true for every choice of i then no equilibrium distribution can satisfy the Detailed Balance Equations.

Example 10.8 (Example 10.1/10.2 revisited). For the Markov chain on page 184, do the Detailed Balance Equations hold?

The Detailed Balance Equation is

$$\pi_1(1/2) = \pi_2(1/4).$$

The equilibrium distribution, $\pi_1 = 1/3$ and $\pi_2 = 2/3$, satisfies the Detailed Balance Equation. This Markov chain is a birth-death chain, so this is the result predicted by Theorem 10.3.

Example 10.9 (Example 10.6 revisited). For the Markov chain on page 189, use the Detailed Balance Equations to find all possible equilibrium distributions.

This Markov chain is a birth-death chain since all possible transitions either change the state by 1 or leave it the same. So we can find all equilibrium distribution(s) by solving the Detailed Balance Equations (with $j = i + 1$ for $i = 0, 1, 2, 3$):

$$\begin{aligned}\pi_0(0) &= \pi_1(1 - p) \\ \pi_1 p &= \pi_2(1 - p) \\ \pi_2 p &= \pi_3(1 - p) \\ \pi_3 p &= \pi_4(0).\end{aligned}$$

Assuming $p \neq 1$, the first three equations give $\pi_1 = 0$, $\pi_2 = 0$, and $\pi_3 = 0$. None of the equations say anything about π_0 and π_4 , so the only constraint is the additional condition $\pi_0 + 0 + 0 + 0 + \pi_4 = 1$. So the equilibrium distributions are given by $\pi_1 = \pi_2 = \pi_3 = 0$, $\pi_4 = 1 - \pi_0$ for all $0 \leq \pi_0 \leq 1$.

This is the same conclusion that we found originally. The algebra is slightly simpler this way – for instance, it is more readily apparent that neither π_0 nor π_4 are constrained in any way, except by the condition $\sum_{i \in S} \pi_i = 1$.

Example 10.10. For the random walk with state space $S = \{0, 1, 2, \dots, N\}$, and transition probabilities

$$\begin{aligned} p_{i,i+1} &= p, & 0 \leq n \leq N-1 \\ p_{i,i-1} &= 1-p=q, & 1 \leq n \leq N \\ p_{00} &= q, & p_{NN} = p \\ p_{ij} &= 0 & \text{otherwise,} \end{aligned}$$

find the equilibrium distribution(s).

This Markov chain is a birth-death chain, so we solve the Detailed Balance Equations:

$$\begin{aligned} \pi_{i-1}p_{i-1,i} &= \pi_i p_{i,i-1}, & \text{for } i > 0 \\ \Rightarrow \pi_{i-1}p &= \pi_i q \\ \Rightarrow \pi_i &= \frac{p}{q} \pi_{i-1} \end{aligned}$$

We express the probabilities π_i in terms of π_0 :

$$\begin{aligned} \pi_1 &= \frac{p}{q} \pi_0 \\ \pi_2 &= \frac{p}{q} \pi_1 = \left(\frac{p}{q}\right)^2 \pi_0 \\ &\vdots \\ \pi_i &= \left(\frac{p}{q}\right)^i \pi_0. \end{aligned}$$

And $\sum_i \pi_i = 1$ implies that

$$\begin{aligned} \pi_0 \left[1 + \frac{p}{q} + \left(\frac{p}{q}\right)^2 + \dots + \left(\frac{p}{q}\right)^N \right] &= 1 \\ \Rightarrow \pi_0 &= \left[1 + \frac{p}{q} + \left(\frac{p}{q}\right)^2 + \dots + \left(\frac{p}{q}\right)^N \right]^{-1} \end{aligned}$$

This can be further simplified by noting that

$$\sum_{i=0}^N \left(\frac{p}{q}\right)^i = \frac{1 - (p/q)^{N+1}}{1 - (p/q)}$$

to give, finally,

$$\pi_i = \frac{(p/q)^i (1 - p/q)}{1 - (p/q)^{N+1}} \quad 0 \leq i \leq N.$$

Example 10.11. For $0 < p < 1$, find the equilibrium distribution(s) for the queueing model with state space $S = \{0, 1, 2, \dots\}$ and transition probabilities

$$\begin{aligned} p_{i,i+1} &= p, \\ p_{i,i-1} &= 1 - p & (i \neq 0) \\ p_{00} &= 1 - p, \\ p_{ij} &= 0 \text{ otherwise.} \end{aligned}$$

This Markov chain is a birth-death chain, so we will solve the Detailed Balance Equations instead of the Full Balance Equations. The only relevant equations are $\pi_i p_{i,i+1} = \pi_{i+1} p_{i+1,i}$ for all $i \in S$, and we note that $p_{i+1,i} = 1 - p$ for all i since $i + 1 \neq 0$. So any equilibrium distribution must satisfy

$$\begin{aligned} p\pi_i &= (1 - p)\pi_{i+1} \\ \pi_{i+1} &= \frac{p}{1 - p}\pi_i \end{aligned}$$

for all $i \geq 0$. By iterating this relation, we obtain

$$\pi_i = \left(\frac{p}{1 - p} \right)^i \pi_0.$$

Putting this into the final condition $\sum_{i \in S} \pi_i = 1$, we obtain

$$1 = \sum_{i=0}^{\infty} \left(\frac{p}{1 - p} \right)^i \pi_0.$$

The infinite series is a geometric series with common ratio $r = p/(1 - p)$. If $r < 1$ (which is equivalent to $p < 1/2$) then the value of the series is a well-defined (finite) number, giving

$$1 = \frac{\pi_0}{1 - r} = \frac{\pi_0}{1 - \frac{p}{1-p}} = \pi_0 \frac{1 - p}{1 - 2p}.$$

In this case we can solve to obtain $\pi_0 = (1 - 2p)/(1 - p)$ and

$$\pi_i = \frac{1 - 2p}{1 - p} \left(\frac{p}{1 - p} \right)^i.$$

Thus exactly one equilibrium distribution exists.

On the other hand, if $p \geq 1/2$, then the common ratio r has $r \geq 1$. In this case the geometric series cannot converge at all unless $\pi_0 = 0$; and if $\pi_0 = 0$ then all terms are 0 and the condition $1 = \sum_{i \in S} \pi_i$ is violated. So if $p \geq 1/2$, no equilibrium distribution can exist.

[This result is (mainly) reasonable from the perspective of the queueing problem. If $p < 1/2$, then at every step the probability is greater that a customer will leave the queue (having finished being served) than the probability that a new customer will join the queue. It is reasonable to suppose that in this case the length of the queue will “equilibrate” in the long run. If $p > 1/2$, then the reverse is true, and more customers are arriving, on average, than leaving. In this case it is reasonable that there should be no equilibrium distribution since customers will continue to pile up. The case $p = 1/2$ (a balance between arrivals and departures, on average) is borderline and the queueing intuition is unclear.]

The meaning of the Full and Detailed Balance Equations

Imagine a large number of particles distributed across the state space of a Markov chain. Suppose the particles are distributed so that the fraction π_i of particles are found at state i at time 0.

At each time step, each particle jumps according to the transition probabilities of the Markov chain, all independently. During the step between time 0 and time 1, particles leave their current state and arrive at a destination state. The fraction of particles that leave state i and enter state j is precisely $\pi_i p_{ij}$ – the fraction π_i of particles were at state i to begin with, and of these a further fraction p_{ij} make the transition to state j . Therefore the Full Balance Equation

$$\pi_j = \sum_{i \in S} \pi_i p_{ij}$$

says that the **total flow out** to state j equals the **total flow in** from all states i . (Particles that return from j to j are counted in both “in” and “out” flows.) If this equation holds (for all j) then the fractions of particles at the various states j will be unchanged at the next step.

By contrast, the Detailed Balance Equations

$$\pi_i p_{ij} = \pi_j p_{ji}$$

say that the **flows between states balance out individually**, not just in total. (The particles emigrating from i to j are balanced by particles immigrating to i from j .) It seems intuitively clear that if all the flows match for each pair of states, then by adding up the flows, the total flows in and out must likewise match.

In fact, this idea underlies the proof of Theorem 10.4, which we now give.

Proof of Theorem 10.4: We analyse the right-hand side in the Full Balance Equations (the total flow in). For any $j \in S$,

$$\begin{aligned} \sum_{i \in S} \pi_i p_{ij} &= \sum_{i \in S} \pi_j p_{ji} && \text{(by the Detailed Balance Equations)} \\ &= \pi_j \sum_{i \in S} p_{ji} \\ &= \pi_j (1) = \pi_j && \text{(since transition probabilities sum to 1)} \end{aligned}$$

Since $j \in S$ can be any state, this completes the proof.

Time reversal

Another interpretation of the Detailed Balance Equations is in terms of *time reversal*. Consider a Markov chain $(X_n, n \geq 0)$, with initial distribution $\mathbb{P}(X_0 = i)$, $i \in S$. Fix a time horizon $N \in \mathbb{N}$ and define

$$Y_n = X_{N-n}, \quad n = 0, 1, \dots, N,$$

the process run backwards in time. The Markov property for $(X_n, n \geq 0)$ operates forwards in time, so it may be surprising to find that the time-reversed process $(Y_n, n = 0, 1, \dots, N)$ *also* satisfies the Markov property. The conditional probabilities $\mathbb{P}(Y_{n+1} = j | Y_n = i)$ depend on the initial distribution $\mathbb{P}(X_0 = i)$, $i \in S$, and for most initial distributions $\mathbb{P}(Y_{n+1} = j | Y_n = i)$ will also depend on n – for instance, if X_0 is a fixed state i_0 , then $\mathbb{P}(Y_N = i_0 | Y_{N-1} = i) = 1$ for any state i for which $\mathbb{P}(Y_{N-1} = i) \neq 0$. However, if the initial distribution is chosen to be an equilibrium distribution $(\pi_i, i \in S)$ that satisfies the Detailed Balance Equations, then the time-reversed process $(Y_n, n = 0, 1, \dots, N)$ is a *time-homogeneous* Markov chain, and the transition probabilities are the *same* as for the original process $(X_n, n \geq 0)$.

The proofs of these statements are left as exercises for the reader.

10.3 Limiting distributions

In Example 10.1 (see page 184), we saw that the distribution of X_n (i.e., the probabilities $\mathbb{P}(X_n = i)$ for $i \in S$) appeared to be converging to fixed values as n increased. We now expand on this notion.

Definition: Consider a Markov chain $(X_n, n \geq 0)$ with initial distribution $\mathbb{P}(X_0 = i), i \in S$. If the limits

$$a_i = \lim_{n \rightarrow \infty} \mathbb{P}(X_n = i)$$

exist for every state $i \in S$, and if

$$\sum_{i \in S} a_i = 1,$$

then $a = (a_i, i \in S)$ is called a *limiting distribution* for the Markov chain.

Example 10.12 (Example 10.1 revisited). Consider the Markov chain defined on page 184.

It is possible to verify that

$$\begin{aligned}\mathbb{P}(X_n = 1) &= \frac{1}{3} \left(1 + \frac{2}{4^n} \right), \\ \mathbb{P}(X_n = 2) &= \frac{2}{3} \left(1 - \frac{1}{4^n} \right).\end{aligned}$$

[We can prove this by mathematical induction. Let \mathcal{S}_n be the statement “the formulas for $\mathbb{P}(X_n = 1)$ and $\mathbb{P}(X_n = 2)$ are correct”. Then \mathcal{S}_0 is correct since we chose the initial condition $\mathbb{P}(X_0 = 1) = 1, \mathbb{P}(X_0 = 2) = 0$, and that is what the formulas say for $n = 0$. Assuming \mathcal{S}_n is correct, we can then prove

$$\begin{aligned}\mathbb{P}(X_{n+1} = 1) &= \mathbb{P}(X_n = 1)p_{11} + \mathbb{P}(X_n = 2)p_{21} \\ &= \frac{1}{2} \cdot \frac{1}{3} \left(1 + \frac{2}{4^n} \right) + \frac{1}{4} \cdot \frac{2}{3} \left(1 - \frac{1}{4^n} \right) \\ &= \frac{1}{2} \cdot \frac{1}{3} \left(1 + \frac{2}{4^n} + 1 - \frac{1}{4^n} \right) \\ &= \frac{1}{2} \cdot \frac{1}{3} \left(2 + \frac{1}{4^n} \right) \\ &= \frac{1}{3} \left(1 + \frac{1}{2 \cdot 4^n} \right) \\ &= \frac{1}{3} \left(1 + \frac{2}{4 \cdot 4^n} \right) = \frac{1}{3} \left(1 + \frac{2}{4^{n+1}} \right)\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}(X_{n+1} = 2) &= 1 - \mathbb{P}(X_{n+1} = 1) \\ &= 1 - \frac{1}{3} \left(1 + \frac{2}{4^{n+1}} \right) \\ &= 1 - \frac{1}{3} - \frac{2}{3} \cdot \frac{1}{4^{n+1}} \\ &= \frac{2}{3} - \frac{2}{3} \cdot \frac{1}{4^{n+1}} = \frac{2}{3} \left(1 - \frac{1}{4^{n+1}} \right).\end{aligned}$$

We have therefore proved \mathcal{S}_{n+1} . By the Principle of Mathematical Induction, the formulas are correct for all $n \geq 0$.

For how these formulas were found in the first place, see Section 10.6 below.]

These formulas confirm our impression that the distribution of X_n was converging. Indeed, since $\lim_{n \rightarrow \infty} 1/4^n = 0$, we have

$$a_1 = \lim_{n \rightarrow \infty} \mathbb{P}(X_n = 1) = \frac{1}{3}, \quad a_2 = \lim_{n \rightarrow \infty} \mathbb{P}(X_n = 2) = \frac{2}{3}.$$

These numbers add up to 1, so $a = (a_1, a_2) = (1/3, 2/3)$ is a limiting distribution for the Markov chain. Moreover this limiting distribution is the same as the equilibrium distribution $\pi = (\pi_1, \pi_2)$ that we found in Section 10.1.

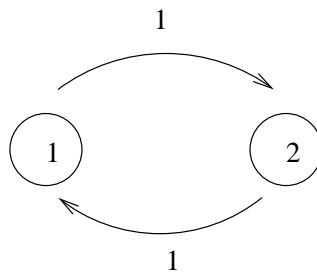
The following theorem says that this is always the case:

Theorem 10.5.

1. If a Markov chain has a limiting distribution $a = (a_i, i \in S)$, then a is also an equilibrium distribution. (That is, setting $\pi_i = a_i$ gives a solution of the Full Balance Equations.)
2. Given an equilibrium distribution $\pi = (\pi_i, i \in S)$, it is always possible to obtain π as a limiting distribution by choosing the initial distribution to be $(\mathbb{P}(X_0 = i) = \pi_i, i \in S)$ (i.e., by starting the Markov chain in equilibrium).

Theorem 10.5 says that the set of possible limiting distributions (over all possible choices of the initial distribution $(\mathbb{P}(X_0 = i), i \in S)$) is the same as the set of equilibrium distributions. However, some initial distributions may have no limiting distribution at all.

Example 10.13. Consider the Markov chain with transition diagram



Start the Markov chain from state 1, i.e., $\mathbb{P}(X_0 = 1) = 1$. Find

- (a) the equilibrium distribution(s); and
- (b) the limiting distribution.

For the equilibrium distribution(s), the Full Balance Equations are

$$\begin{aligned}\pi_1 &= \pi_2 \\ \pi_2 &= \pi_1\end{aligned}$$

and the condition $\pi_1 + \pi_2 = 1$ then gives the unique solution

$$\pi_1 = \pi_2 = \frac{1}{2}.$$

For the limiting distribution, we will find $\mathbb{P}(X_n = 1)$ (and $\mathbb{P}(X_n = 2) = 1 - \mathbb{P}(X_n = 1)$) for all $n \geq 0$. In this example, there is no randomness and the Markov chain always alternates back and forth between states 1 and 2. Therefore

$$\mathbb{P}(X_n = 1) = \begin{cases} 1 & \text{if } n \text{ is even,} \\ 0 & \text{if } n \text{ is odd.} \end{cases}$$

The sequence $1, 0, 1, 0, 1, \dots$ has *no limit*, and so *no limiting distribution exists* for this initial distribution.

Thus in this example, an equilibrium distribution exists even though there is no limiting distribution. We say this chain is *periodic* with period 2, because we can only return to a state at even steps. A chain that is periodic need not have a limiting distribution, even though it may have an equilibrium distribution.

Example 10.14 (Example 10.7 revisited). Let $p > 0$. Find the limiting distribution for the Markov chain from page 190, with state space $S = \{0, 1, 2, \dots\}$, initial state $X_0 = 0$, and transitions described by

$$X_{n+1} = \begin{cases} X_n + 1 & \text{with probability } p, \\ X_n & \text{with probability } 1 - p. \end{cases}$$

Use the following fact:

$$\lim_{n \rightarrow \infty} n^i x^{n-i} = 0 \quad \text{for any fixed values } i \in \mathbb{N}, -1 < x < 1.$$

We would like to find the distribution of X_n . In this example, we claim that X_n will have the Binomial(n, p) distribution. To see this, define the “increment” variables

$$I_n = X_n - X_{n-1},$$

so that $X_n = X_0 + I_1 + I_2 + \dots + I_n$. (Writing out the definition of I_1, I_2, \dots, I_n , the sum “telescopes” and only X_n remains.) Since $X_0 = 0$, X_n counts the number of I_i ’s that are 1. Furthermore I_1, \dots, I_n form a sequence of Bernoulli trials with success probability p , where “success” means the value 1 and “failure” means 0. (For instance, $\mathbb{P}(I_1 = 1, I_2 = 0, I_3 = 1) = \mathbb{P}(X_1 = 1, X_2 = 1, X_3 = 2) = p_{01}p_{11}p_{12} = p(1-p)p$, and similarly for any sequence of values for I_1, \dots, I_n .)

Since X_n has the Binomial(n, p) distribution,

$$\begin{aligned} \mathbb{P}(X_n = i) &= \binom{n}{i} p^i (1-p)^{n-i} \\ &= \frac{n!}{(n-i)!} (1-p)^{n-i} \frac{p^i}{i!} \\ &= n(n-1)(n-2) \cdots (n-i+1) (1-p)^{n-i} \frac{p^i}{i!} \\ &\leq n^i (1-p)^{n-i} \frac{p^i}{i!}. \end{aligned}$$

The quantity on the right-hand side approaches 0 as $n \rightarrow \infty$ since $0 \leq 1-p < 1$. Since $\mathbb{P}(X_n = i)$ cannot be smaller than 0, we must have $\lim_{n \rightarrow \infty} \mathbb{P}(X_n = i) = 0$ also.

[This is the “Squeeze Theorem” for limits.]

Thus $a_i = \lim_{n \rightarrow \infty} \mathbb{P}(X_n = i) = 0$ for every i . The condition $\sum_{i \in S} a_i = 1$ is violated, so $(a_i = 0, i \in S)$ is *not* a limiting distribution.

It is possible to adapt this argument when the initial distribution is not $X_0 = 1$. In all cases, the conclusion is that *no limiting distribution exists*.

We can see this another way without having to do any calculation. We already showed that no *equilibrium* distribution exists (see page 190). Part 1 of Theorem 10.5 says that any limiting distribution would also have to be an equilibrium distribution as well. Since no equilibrium distribution exists, no limiting distribution can exist either.

10.4 Qualitative features of Markov chains

In the examples so far, we have seen that a Markov chain may have a unique equilibrium distribution, infinitely many equilibrium distributions, or no equilibrium distributions. Likewise, we know that any limiting distribution, if it exists, must be an equilibrium distribution (Theorem 10.5), but even when an equilibrium distribution exists there may be no limiting distribution. We would like to have general conditions to tell us when these various possibilities may occur, without having to make calculations – especially for limiting distributions, where it is generally quite difficult to calculate the probabilities $\mathbb{P}(X_n = i)$ directly.

Definition: A state $i \in S$ is called *absorbing* if $p_{ii} = 1$.

An absorbing state is a state from which the Markov chain never leaves.

Definition: We say that you can reach from state i to state j , and we write $i \rightsquigarrow j$, if there is a path $i \rightarrow s_1 \rightarrow s_2 \rightarrow \cdots \rightarrow s_{n-1} \rightarrow j$ (of arbitrary length) such that the transition probabilities $p_{is_1}, p_{s_1s_2}, \dots, p_{s_{n-2}s_{n-1}}, p_{s_{n-1}j}$ are all non-zero. By convention, we say you can reach from any state to itself, i.e., $i \rightsquigarrow i$.

We say that i and j *communicate* if $i \rightsquigarrow j$ and $j \rightsquigarrow i$.

Definition: A Markov chain is called *irreducible* if all states communicate.

Definition: An irreducible Markov chain is called *periodic* if it is possible to partition the state space S into k disjoint sets S_1, S_2, \dots, S_k , $k \geq 2$, such that every transition from every state in S_i goes to some state in S_{i+1} ($i < k$) and every transition from every state in S_k goes to some state in S_1 .

An irreducible Markov chain that is not periodic is called *aperiodic*.

By convention, we only apply the periodic/aperiodic distinction to irreducible Markov chains.

All of these properties depend only on the set of possible transitions (i.e., transitions with non-zero transition probability). Visually, they depend on the arrows in a transition diagram, but not on the probabilities associated to the arrows (so long as they are non-zero). Several shortcuts are available in common situations:

- If a Markov chain has any absorbing states, then it is *not* irreducible.⁶
- If there is a loop $i \rightarrow s_1 \rightarrow \cdots \rightarrow s_{n-1} \rightarrow i$ that visits every state and then returns to the start (with every arrow corresponding to a possible transition, possibly with repeated visits to the same state) then the Markov chain is irreducible.

⁶Except for Markov chains with only one state.

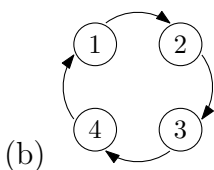
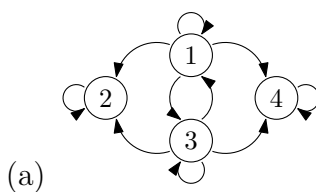
- If the Markov chain is *not* irreducible, then it is possible to find at least one subset $S' \subset S$ (neither empty nor containing all states) such that all transitions from all states in S' goes to states that are still in S' .

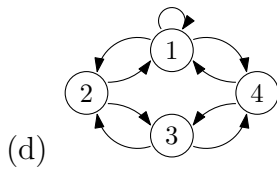
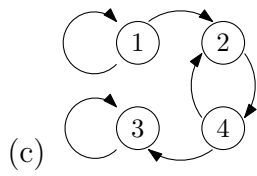
And, assuming we are talking about an irreducible Markov chain,

- If even one state i has $p_{ii} > 0$, then the Markov chain is aperiodic.
- If the Markov chain always jumps between even-numbered states and odd-numbered states (with no possibility of remaining at the same state) then the Markov chain is periodic. *[This is the most common kind of periodicity, and the only kind of periodicity that birth-death chains can have.]*
- Consider all numbers $n \in \mathbb{N}$ for which there is a loop in the Markov chain of length exactly n (i.e., consisting of exactly n transitions, including the last one). Call p the greatest common divisor (greatest common factor) of all of these numbers. If $p \neq 1$ then the Markov chain is periodic (and it can be partitioned as in the definition of periodicity with $k = p$).
- If the Markov chain contains a loop of length n and a loop of length m , and if n and m have no common divisors, then the Markov chain is aperiodic.

Example 10.15. For each of the following Markov chains, determine

- whether or not the chain has any absorbing state(s);
- whether the chain is irreducible or reducible; and
- (if the chain is irreducible) whether the chain is periodic or aperiodic.





Some other interesting facts left as exercises for the reader:

- If i is an absorbing state then $\pi_i = 1$, $\pi_j = 0$ for $j \neq i$ is an equilibrium distribution.
- If $(\pi_i, i \in S)$ is an equilibrium distribution for an irreducible Markov chain, then $\pi_i > 0$ for all $i \in S$.
- If $(\pi_i, i \in S)$ is an equilibrium distribution for a periodic Markov chain, and if the state space has been partitioned into S_1, S_2, \dots, S_k as in the definition of periodicity, then

$$\sum_{i \in S_1} \pi_i = \sum_{i \in S_2} \pi_i = \dots = \sum_{i \in S_k} \pi_i = \frac{1}{k}.$$

10.5 When are the equilibrium and limiting distributions equal?

The following theorem gives a condition under which the limiting distribution is *guaranteed* to equal the equilibrium distribution.

Theorem 10.6. If a Markov chain $\{X_n\}_{n \geq 0}$ with state space S

- (a) has an equilibrium distribution $\{\pi_i\}_{i \in S}$, and
- (b) is irreducible (i.e. all states communicate), and
- (c) is aperiodic,

then the limiting distribution exists and is equal to the equilibrium distribution $\{\pi_i\}_{i \in S}$.

In Theorem 10.6 above, the limiting distribution does not depend on the initial distribution, however the speed of convergence to the limiting distribution does depend on the initial distribution. For example, if the conditions of the theorem are satisfied and the initial distribution is given by the equilibrium distribution, i.e. $\mathbb{P}(X_0 = i) = \pi_i$ for each i , then the distribution is already equal to the limiting distribution!

Corollary: If a Markov chain is irreducible and aperiodic, then it can have at most one equilibrium distribution.

This follows from Theorem 10.6: if the Markov chain has an equilibrium distribution π , then any limiting distribution must be the same as π . There cannot be another (different) equilibrium distribution π' , for if there were then we could take π' as the initial distribution to obtain π' as the limiting distribution, contradicting Theorem 10.6.

[In fact the corollary is true for all irreducible Markov chains, whether periodic or aperiodic. This does not follow directly from Theorem 10.6, although it is possible to use Theorem 10.6 as part of a proof of this stronger statement.]

If any of the conditions in Theorem 10.6 are false, the conclusion may not be true. The most striking example is the following:

Theorem 10.7. If a Markov chain is irreducible and periodic, and if it is started from a fixed state (i.e., $\mathbb{P}(X_0 = i) = 1$ for some state $i \in S$), then the limiting distribution *does not exist*.

The proof of Theorem 10.7 follows similar reasoning to Example 10.13: each sequence of probabilities $\mathbb{P}(X_n = j)$ must contain infinitely many zeroes, so it either has no limit or has limit 0. Hence the numbers a_j either do not exist or else exist and are all 0, and in either case they do not give a limiting distribution.

10.6 Optional section: Equilibrium distributions, limiting distributions, and transition matrices

[This section assumes familiarity with linear algebra and matrices. If you are not familiar with these topics, you should omit this section.]

For Markov chains with finite state spaces, we can use linear algebra and the theory of matrices to understand equilibrium distributions and limiting distributions. We suppose that the state space is $S = \{1, 2, \dots, n\}$ or $S = \{0, 1, \dots, n-1\}$, so that, in either case, the transition matrix is $n \times n$. Equilibrium distributions are *row vectors*, i.e., $1 \times n$ matrices:

$$\pi = (\pi_1 \quad \pi_2 \quad \cdots \quad \pi_n) \quad \text{or} \quad \pi = (\pi_0 \quad \pi_1 \quad \cdots \quad \pi_{n-1}).$$

(We omit the commas for consistency with matrix notation.) That is, the j^{th} entry of π (in the first row and j^{th} column) is π_j .

In terms of matrices, the Full Balance Equations say

$$\pi P = \pi,$$

plus the additional condition

$$\pi \mathbb{1} = 1,$$

where $\mathbb{1}$ is the column vector (of length n) with every entry being 1. Alternatively we can write

$$\pi(I - P) = 0 \quad \text{or} \quad (I - P^T)\pi^T = 0$$

where I is the $n \times n$ identity matrix and T denotes transpose (so that π^T is a column vector). The matrix $(I - P)$ is always singular: in fact its nullspace contains the vector $\mathbb{1}$ since $P\mathbb{1} = \mathbb{1}$. Hence the transpose matrix $(I - P)^T$ is also singular, so there is always at least one non-zero solution v to the equation $v(I - P) = 0$. It turns out to be always possible to arrange for v to have non-negative entries (not all 0), and by rescaling this v we can always obtain an equilibrium distribution. (This is a way to prove Theorem 10.2.)

Note: In linear algebra, it is usually preferred to have *column* vectors as unknowns. This is often arranged by replacing the transition matrix P by the matrix with p_{ij} in the i^{th} column and j^{th} row (the *transpose* of our transition matrix). As a result, some texts write the above equations differently.

If you wish to solve the Full Balance Equations using matrices, you should solve the equation $(I - P^T)\pi^T = 0$ for the column vector π^T . *Do not* attempt to solve $\pi(I - P) = 0$ using standard Gaussian elimination or augmented matrices, as these do not apply when the unknown is a row vector on the left of the matrix.

For a Markov chain with a specific initial distribution, let q_n denote the row vector with entries $\mathbb{P}(X_n = j)$. (Thus the initial distribution is specified by q_0 .) The analogue of Equation (*) is

$$q_{n+1} = q_n P, \quad q_n = q_0 P^n.$$

The behaviour of P^n for $n \rightarrow \infty$ is determined by the eigenvalues of P . We can write $P = ADA^{-1}$, where D is the diagonal matrix of eigenvalues (or, in some cases, a Jordan block

matrix) and A is an invertible matrix of eigenvectors. For a transition matrix, 1 is always an eigenvalue since $(I - P)$ is singular. The eigenvectors for $\lambda = 1$ are the span of the equilibrium distributions. If the Markov chain is irreducible, there is only one equilibrium distribution and the eigenvalue $\lambda = 1$ is simple, i.e., it is not a repeated eigenvalue. The other eigenvalues can in general be complex numbers, but must always satisfy $|\lambda| \leq 1$.

Taking powers,

$$P^n = AD^nA^{-1}.$$

If all other eigenvalues satisfy $|\lambda| < 1$, then the matrix D^n will converge as $n \rightarrow \infty$ to a diagonal matrix with all entries either 0 or 1. When in addition the Markov chain is irreducible, only one diagonal entry is non-zero, and it follows that P^n converges to a matrix where every row is equal to the unique equilibrium distribution π .

If there is another eigenvalue with $|\lambda| = 1$, $\lambda \neq 1$, then λ must be a root of unity, $\lambda^r = 1$ for some $r \in \{2, 3, \dots\}$. (If the Markov chain is actually irreducible, what we have called periodicity is equivalent to some eigenvalue $\lambda \neq 1$ being a root of unity.) Hence the sequence of powers $\lambda, \lambda^2, \lambda^3, \dots$ must repeat periodically, but is not constant; it follows that the limiting distribution may not exist for certain initial distributions.

By finding a decomposition $P = ADA^{-1}$ (or $P = B^{-1}DB$) it is possible to explicitly calculate n -step transition probabilities and distributions of X_n . This is what was done in Example 10.1 on page 197. However, this procedure can be difficult if the number of states is large, and may not apply at all when the state space is infinite.

Question: In this matrix perspective, distributions (equilibrium distributions π and the distributions q_n for X_n) correspond to row vectors, with a time step corresponding to multiplication on the right by P . What do column vectors (with a time step corresponding to multiplication on the left by P) correspond to?

10.7 Exercises

10.7.1 A Markov chain with state-space $S = \{0, 1, 2\}$ has the following transition matrix,

$$\begin{pmatrix} 0 & 1 & 0 \\ 0.5 & 0 & 0.5 \\ 0 & 1 & 0 \end{pmatrix}$$

1. Find $p_{00}^{(4)}$, $p_{01}^{(4)}$ and $p_{02}^{(4)}$.
2. Does the equilibrium distribution exist? If so, find it. If not, explain why.
3. Suppose that $\mathbb{P}(X_0 = i) = 1$ for some $i \in S$. Does the limiting distribution exist? If so, find it. If not, explain why not.

10.7.2 A Markov chain with state-space $S = \{0, 1, 2\}$ has the following transition matrix,

$$\begin{pmatrix} 0.1 & 0.9 & 0 \\ 0.5 & 0 & 0.5 \\ 0 & 1 & 0 \end{pmatrix}$$

1. Find $p_{00}^{(4)}$, $p_{01}^{(4)}$ and $p_{02}^{(4)}$.
2. Does the equilibrium distribution exist? If so, find it. If not, explain why not.
3. Does the limiting distribution exist? If so, find it. If not, explain why not.

10.7.3 For the asymmetric simple random walk with right-step probability p , find a formula for $p_{0m}^{(n)}$.

10.7.4 A Markov chain with two states, labelled 1 and 2, has the following transition matrix:

$$P = \begin{pmatrix} 1/4 & 3/4 \\ 2/3 & 1/3 \end{pmatrix}.$$

1. Write down the Full Balance Equations for this Markov chain.
2. Obtain the equilibrium distribution for this Markov chain.

10.7.5 Consider a Markov chain with states $S = \{1, 2, 3\}$ and transition matrix:

$$\begin{pmatrix} 0 & 2/3 & 1/3 \\ 1/3 & 0 & 2/3 \\ 1/3 & 2/3 & 0 \end{pmatrix}$$

1. Draw the transition diagram for this Markov chain.
2. Write down the equilibrium equations for this Markov chain.
3. Obtain the equilibrium distribution for this Markov chain.

10.7.6 A total of M balls are divided between two containers. At each time point, one ball is chosen at random from the M balls (each ball is equally likely to be chosen), and that ball is transferred from the container it is in to the other container (so if the selected ball is in container 1 it is transferred to container 2, and if it is in container 2, it is transferred to container 1). Let X_n denote the number of balls in container 1 at time n .

1. Start by supposing $M = 3$. Write down the Detailed Balance Equations that the equilibrium distribution satisfies for this chain, and hence, or otherwise, find the equilibrium distribution for this chain (you must show working here).
2. Now consider the chain with general M . As before, write down the Detailed Balance Equations that the equilibrium distribution satisfies for this chain. Show, by substitution, that the equilibrium distribution is given by

$$\pi_i = \binom{M}{i} \left(\frac{1}{2}\right)^M$$

for $i = 0, 1, 2, \dots, M$. What is this distribution? Give an explanation of why this should be the equilibrium distribution.

10.7.7 Let $\{X_n, n \geq 0\}$ be a Markov chain with state space $S = \{1, 2, 3\}$, and transition matrix

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{pmatrix}.$$

1. Write down the equilibrium equations for this Markov chain.
2. Obtain the equilibrium distribution for this Markov chain.

10.7.8 Consider the Markov chain $\{X_n; n = 0, 1, 2, \dots\}$ with state space $S = \{0, 1, 2\}$ and transition matrix

$$P = \begin{pmatrix} 0 & 1 & 0 \\ 1/3 & 1/3 & 1/3 \\ 0 & 1/3 & 2/3 \end{pmatrix}.$$

1. Write down the equations that the equilibrium distribution satisfies. Please give both
 - (a) the Full Balance Equations and
 - (b) the Detailed Balance Equations (note that this is a birth and death chain).
2. Find the equilibrium distribution for this Markov chain.

10.7.9 A professor gives tests that are hard, medium, or easy. If she gives a hard test, her next test will be either medium or easy, with equal probability. However, if she gives a medium or easy test, there is a 0.5 probability that her next test will be of the same difficulty, and a 0.25 probability for each of the other two levels of difficulty. Construct an appropriate Markov chain and find the probability in equilibrium that she gives an easy test.

10.7.10 A lecturer has two umbrellas that she uses when commuting from home to the office and back. If it is raining as she is about to leave her current location, and an umbrella is available there, then she takes it with her. If it is not raining, she doesn't take an umbrella with her. Suppose that it rains with probability p each time she commutes, independently of all the other times.

1. Show that the equilibrium probability that she gets wet during a commute is $p \times \frac{(1-p)}{(3-p)}$.
Hint: Let X_n be the number of umbrellas at her current location after n journeys.
2. Find the value of p that maximises the equilibrium probability of getting wet.

10.7.11 A total of two fleas live on N dogs. Both of the fleas jump at the same time, but otherwise behave independently of each other. When a flea jumps, it either jumps onto a different location on the same dog (with probability p), or onto any one of the other dogs with equal probability, independently of previous jumps. Let X_n be the number of fleas on dog 1 after n jump times have occurred.

1. What is the state space of the Markov chain $\{X_n\}_{n \geq 0}$?
2. What is the equilibrium distribution for this Markov chain?
3. What is the limiting distribution for this Markov chain?
4. Give the appropriate transition matrix P for this Markov chain.

10.7.12 Two yachts, sailed by “Alonghi” and “Team New Wailand” respectively are sailing around a course. If the teams are even at the beginning of a lap then during that lap they embark on a duel and one team gains a 1 boat length advantage by the end of the lap. Otherwise the leading yacht always sails according to the current wind, while the other yacht sails separately, hoping for a wind-shift. This means that on each lap of the course the leading yacht gains one boat length over the other yacht if there is no wind-shift, while the team behind gains one boat length if the wind-shifts on that lap. Suppose that Team New Wailand wins any given duel with probability $p_1 \in (0, 1)$ (independent of all previous duels and wind-shifts) and that with probability $p_2 \in (0, 1)$ there is a wind-shift on any given lap (independent of all previous wind-shifts and duels).

1. Find the probability that Team New Wailand is leading the race at the end of the first lap if:

- (a) The teams are even at the beginning of the race.
- (b) The winner of a pre-race duel starts with a 1 boat length lead.

Hereafter assuming that there is a pre-race duel, let X_n be the number of boat lengths that Team New Wailand leads by at the end of n laps.

2. Draw the transition diagram for the Markov chain X_n .
3. Give the probability function for X_0 .
4. Find the probability that Team New Wailand is leading the race at the end of 2 laps.
5. *For what values of p_2 would this Markov chain have an equilibrium distribution?
6. *Show that the probability that the race is tied at the end of 3 laps is increasing in p_2 .

11 Hitting/reaching probabilities and times

By the end of this chapter you should be able to:

- set up and solve the system of equations to compute hitting probabilities for states of a Markov chain
- compute expected hitting times based on the transition matrix

Two general questions of interest are

“If the chain is in state i , what is the probability that it will reach state j ?”

“If the chain is in state i , what is the expected time until it reaches state j ?”

Example 11.1. a) For an insurance company what is the probability of ruin (in the next year)?

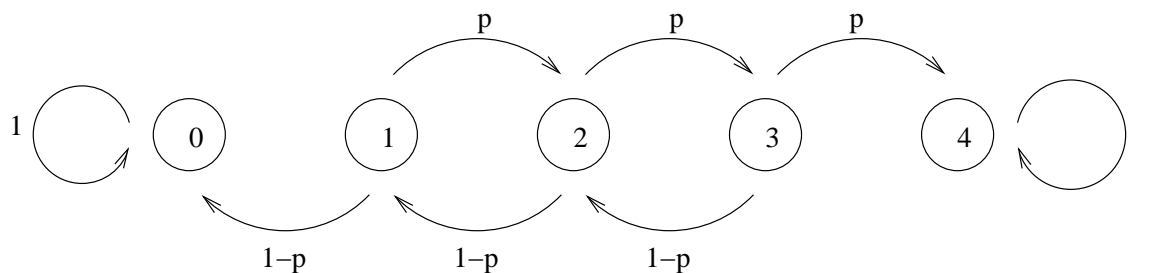
- b) For a population of native birds (e.g. kiwi) what is the probability that they will become extinct?
- c) What is the expected time until you finish your degree (assuming that you pass every course with probability p , and do 4 courses every semester)?
- d) What is the probability a gambler wins \$400 before losing all his/her money? How long will the gambler get to play, on average, before one of these two outcomes occurs?

11.1 Hitting probabilities

Let's start with gambler's ruin problem.

Starting from \$200, what is the probability a gambler reaches \$400 for the symmetric random walk with $p = \frac{1}{2}$, betting \$100 every time?

Let state i denote $i \times \$100$.



Here $p = \frac{1}{2}$.

Let

$$h_{ij} = \mathbb{P}(X_n = j \text{ for some } n \geq 0 | X_0 = i)$$

be the probability of reaching j , starting from i .

We wish to find h_{24} . Start by conditioning on the first step. With probability $\frac{1}{2}$, a Head is thrown, and the chain moves to state 3. From there, by the Markov property, the probability of reaching state 4 is h_{34} . W. p. $\frac{1}{2}$, a Tail is thrown and the chain moves to state 1. From there, by the Markov property, the probability of reaching state 4 is h_{14} . Thus we have

$$h_{24} = p_{23}h_{34} + p_{21}h_{14} = \frac{1}{2}h_{34} + \frac{1}{2}h_{14}.$$

Now we need to find h_{34} and h_{14} .

h_{34} first. With probability $\frac{1}{2}$, a Head is thrown and the chain moves to state 4 — it's reached \$400! (Note that $h_{44} = 1$). With probability $\frac{1}{2}$, a Tail is thrown and the chain returns to state 2. So

$$h_{34} = p_{34}h_{44} + p_{32}h_{24} = \frac{1}{2} + \frac{1}{2}h_{24}.$$

Finally,

$$h_{14} = p_{10}h_{04} + p_{12}h_{24}$$

and $h_{04} = 0$ so

$$h_{14} = \frac{1}{2}h_{24}$$

Thus

$$\begin{aligned} h_{24} &= \frac{1}{2} \left(\frac{1}{2} + \frac{1}{2}h_{24} \right) + \frac{1}{2} \left(\frac{1}{2}h_{24} \right) \\ &= \frac{1}{4} + \frac{1}{2}h_{24} \\ \Rightarrow \frac{1}{2}h_{24} &= \frac{1}{4} \\ \Rightarrow h_{24} &= \frac{1}{2}. \end{aligned}$$

Why does this not surprise us? By symmetry, $h_{24} = h_{20} = \frac{1}{2}$.

Theorem 11.1. The general equations for reaching (hitting) probabilities are

$$h_{ij} = \sum_{k \in S} p_{ik} h_{kj} \quad i \neq j \in S,$$

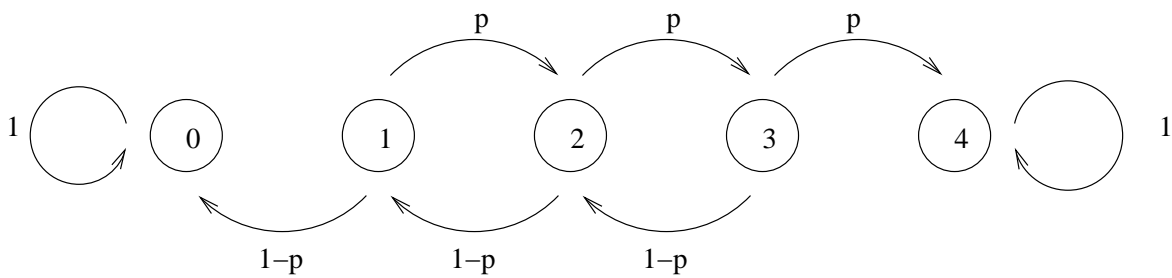
$$h_{jj} = 1.$$

Furthermore

$$h_{ij} = 0$$

for any state i from which it is impossible to reach j , including the case where i is any absorbing state other than j .

Example 11.2. What is h_{24} for the random walk with $S = \{0, 1, 2, 3, 4\}$ and absorbing barriers at states 0 and 4, when $p \neq 1/2$?



$$\begin{aligned} h_{24} &= ph_{34} + (1-p)h_{14} \\ h_{34} &= p + (1-p)h_{24} \\ h_{14} &= ph_{24} + (1-p)h_{04} \\ &\quad \text{(note that } h_{04} = 0) \end{aligned}$$

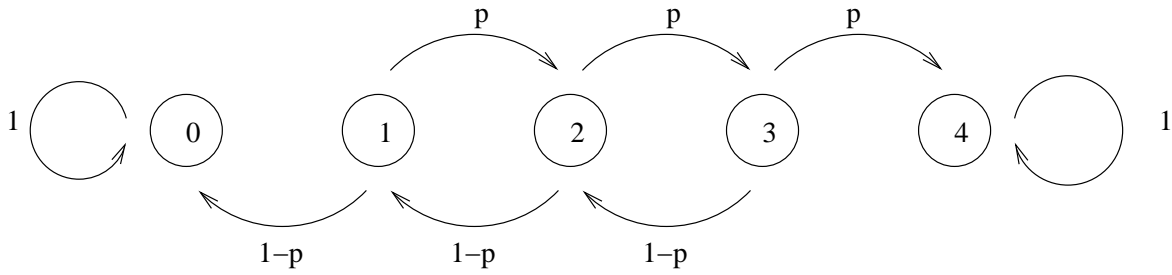
$$\begin{aligned} h_{24} &= p(p + (1-p)h_{24}) + (1-p)ph_{24} \\ &= p^2 + 2p(1-p)h_{24} \\ \Rightarrow h_{24}(1 - 2p(1-p)) &= p^2 \\ \Rightarrow h_{24} &= \frac{p^2}{1 - 2p(1-p)} \end{aligned}$$

11.2 Expected hitting times

In this section we consider the expected time (time here is the number of steps) until we reach a set of states for the first time.

Example 11.3. Gambler's ruin with $p = 1/2$ (continued).

What is the expected time until we reach states 0 or 4, starting from state 2?



Here $p = \frac{1}{2}$.

Let m_{iA} be the expected number of steps (time) to reach the set A , starting from state i at time 0.

We wish to find m_{2A} = expected time to reach $A = \{0, 4\}$.

Starting from state 2, the chain takes 1 step, and then, w. p. $\frac{1}{2}$ moves to state 3, and w. p. $\frac{1}{2}$ moves to state 1.

$$\begin{aligned}
 \text{So } m_{2A} &= 1 + \frac{1}{2} m_{3A} + \frac{1}{2} m_{1A} \\
 m_{3A} &= 1 + \frac{1}{2} m_{4A} + \frac{1}{2} m_{2A} = 1 + \frac{1}{2} m_{2A}, & \text{since } m_{4A} = 0. \\
 m_{1A} &= 1 + \frac{1}{2} m_{2A} + \frac{1}{2} m_{0A} = 1 + \frac{1}{2} m_{2A}, & \text{since } m_{0A} = 0. \\
 & (= m_{3A}, \text{ by observation}).
 \end{aligned}$$

$$\begin{aligned}
 \text{Hence } m_{2A} &= 1 + m_{3A} = 1 + 1 + \frac{1}{2} m_{2A} \\
 \Rightarrow m_{2A} &= 2 + \frac{1}{2} m_{2A} \\
 \Rightarrow \frac{1}{2} m_{2A} &= 2 \quad \Rightarrow m_{2A} = 4 \text{ steps.}
 \end{aligned}$$

And we also have $m_{1A} = m_{3A} = 3$.

Theorem 11.2. The expected hitting times m_{iA} satisfy

$$\begin{aligned} m_{iA} &= 1 + \sum_{k \in S} p_{ik} m_{kA} & i \notin A, \\ m_{iA} &= 0 & i \in A. \end{aligned}$$

Example 11.4. What is m_{2A} in the same Gambler's ruin problem when $p \neq 1/2$?

$$\begin{aligned} m_{2A} &= 1 + pm_{3A} + (1-p)m_{1A} \\ m_{3A} &= 1 + pm_{4A} + (1-p)m_{2A} \\ m_{1A} &= 1 + pm_{2A} + (1-p)m_{0A} \\ &\text{with } m_{0A} = m_{4A} = 0 \end{aligned}$$

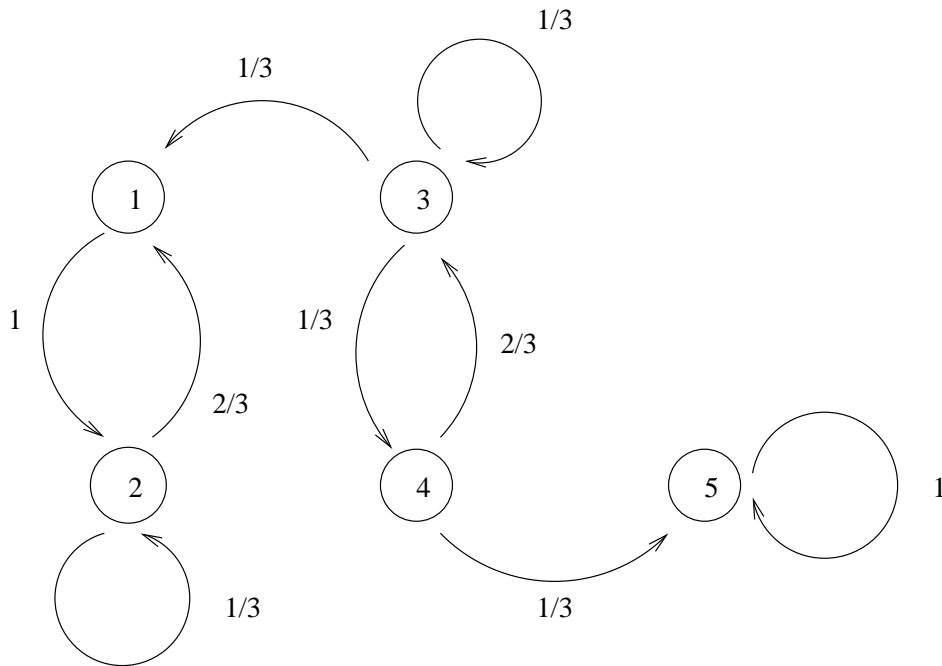
$$\begin{aligned} \Rightarrow m_{2A} &= 1 + p + p(1-p)m_{2A} \\ &\quad + (1-p) + p(1-p)m_{2A} = 2 + 2p(1-p)m_{2A} \end{aligned}$$

$$\Rightarrow m_{2A} = (1 - 2p(1-p)) = 2 \quad \Rightarrow \quad m_{2A} = \frac{2}{1 - 2p(1-p)}$$

Example 11.5. Let $\{X_n, n \geq 0\}$ be a Markov chain with states $\{1, 2, 3, 4, 5\}$ and the following transition matrix:

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 2/3 & 1/3 & 0 & 0 & 0 \\ 1/3 & 0 & 1/3 & 1/3 & 0 \\ 0 & 0 & 2/3 & 0 & 1/3 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

1. Draw the transition diagram for this chain.



2. What is $h_{45} = \mathbb{P}(X_n = 5 \text{ for some } n \geq 0 | X_0 = 4)$, i.e. the probability that the chain ever reaches state 5, given that it started in state 4 at time 0?

$$h_{45} = \frac{1}{3}h_{55} + \frac{2}{3}h_{35} = \frac{1}{3} + \frac{2}{3}h_{35} \quad \text{since } h_{55} = 1$$

$$h_{35} = \frac{1}{3}h_{15} + \frac{1}{3}h_{35} + \frac{1}{3}h_{45} = \frac{1}{3}h_{35} + \frac{1}{3}h_{45} \quad \text{since } h_{15} = 0$$

$$\Rightarrow \frac{2}{3}h_{35} = \frac{1}{3}h_{45} \Rightarrow h_{35} = \frac{1}{2}h_{45}$$

$$h_{45} = \frac{1}{3} + \frac{2}{3} \times \frac{1}{2}h_{45} \Rightarrow \frac{2}{3}h_{45} = \frac{1}{3} \Rightarrow h_{45} = \frac{1}{2}$$

3. What is h_{42} ?

$$h_{42} = 1 - h_{45} = \frac{1}{2}.$$

4. Let $A = \{1, 2, 5\}$. What is m_{3A} ?

$$\begin{aligned} m_{3A} &= 1 + \frac{1}{3}m_{1A} + \frac{1}{3}m_{3A} + \frac{1}{3}m_{4A} \\ m_{4A} &= 1 + \frac{1}{3}m_{5A} = \frac{2}{3}m_{3A} \\ m_{1A} &= 0 = m_{5A} \end{aligned}$$

Hence we obtain

$$\begin{aligned} m_{3A} &= 1 + \frac{1}{3}m_{1A} + \frac{1}{3} + \frac{2}{9}m_{3A} \\ \Rightarrow m_{3A} &= \frac{4}{3} + \frac{5}{9}m_{3A} \\ \Rightarrow \frac{4}{9}m_{3A} &= \frac{4}{3} \\ \Rightarrow m_{3A} &= 3. \end{aligned}$$

Summary

Definitions

- $h_{ij} = \mathbb{P}(X_n = j \text{ for some } n \geq 0 | X_0 = i)$
is the probability that the Markov chain ever reaches state j given that it started in state i .
- m_{iA} is the expected time for the Markov chain to reach the set A , given that it started in state i .

$$h_{ij} = \sum_{k \in S} p_{ik} h_{kj} \quad i \neq j \in S.$$

$$h_{ii} = 1.$$

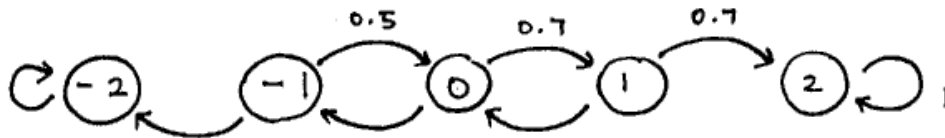
and

$$m_{iA} = 1 + \sum_{k \in S} p_{ik} m_{kA} \quad i \notin A$$

$$m_{iA} = 0 \quad i \in A.$$

11.3 Exercises

- 11.3.1 In a game of tennis, one player is the server, while the other is the receiver. Sometimes the situation “deuce” arises whereby the scores are tied and either player just needs to get two points ahead of the other in order to win the game. The server usually has the advantage, but gets nervous when behind by one point. Let X_n be the difference between the number of points of the server and the receiver, n points after the deuce situation first arises. Suppose that X_n is a Markov chain with transition diagram of the form:



1. Give the missing transition probabilities.
 2. Find $p_{02}^{(2)}$ and $p_{0-2}^{(2)}$ and interpret these quantities in terms of the tennis game.
 3. Find the probability that the server wins the game, starting from the situation “deuce”.
 4. Does the limiting distribution exist? If so, find it. If not, explain why.
- 11.3.2 Let $\{X_n, n \geq 0\}$ be a Markov chain with states $\{1, 2, 3, 4, 5, 6, 7, 8\}$ and the following transition matrix:

$$P = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{matrix} \\ \begin{pmatrix} 1/4 & 1/4 & 1/4 & 0 & 1/4 & 0 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 & 0 & 0 & 0 & 0 \\ 0 & 1/4 & 0 & 0 & 0 & 1/4 & 1/4 & 1/4 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{matrix}$$

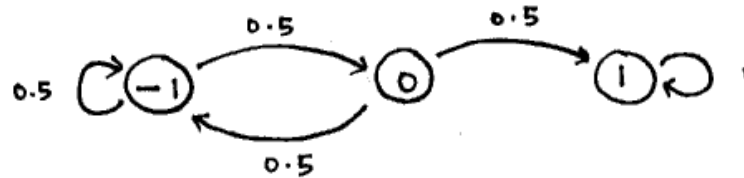
- (a) Draw the transition diagram for this chain.
- (b) Suppose the chain starts in state 1.
 - (i) What is the probability that state 2 is ever reached?
 - (ii) What is the probability that state 8 is ever reached?

11.3.3 Consider the Markov chain $\{X_n; n = 0, 1, 2, \dots\}$ with state space $S = \{0, 1, 2, 3\}$ and transition matrix

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 \\ 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

1. What is h_{23} ?
2. What is h_{20} ?

11.3.4 A Markov chain has the following transition diagram:



1. Find the limiting distribution, i.e. $\lim_{n \rightarrow \infty} \mathbb{P}(X_n = i)$, for $i \in S$.
2. Find the expected time to reach state 1, starting from state -1 .

11.3.5 A Markov chain with five states, labelled $1, 2, 3, 4, 5$, has the following transition matrix:

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1/3 & 0 & 1/3 & 1/3 & 0 \\ 0 & 1/2 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 2/3 & 1/3 \\ 0 & 0 & 0 & 1/2 & 1/2 \end{pmatrix}.$$

1. Draw the transition diagram for this Markov chain.
2. Find $h_{34} = P(X_n = 4 \text{ for some } n \geq 0 | X_0 = 3)$.
3. Find $h_{31} = P(X_n = 1 \text{ for some } n \geq 0 | X_0 = 3)$.
4. Let $A = \{1, 4, 5\}$. Find m_{3A} , the expected number of steps to reach the set of states A , given that the process starts in state 3.

11.3.6 Let $\{X_n\}$ be a Markov chain with state-space $S = \{1, 2, 3, 4, 5, 6\}$ and the following transition matrix:

$$P = \begin{pmatrix} 1/2 & 0 & 0 & 1/2 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 1/2 & 0 \\ 1/3 & 1/3 & 0 & 0 & 0 & 1/3 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/3 & 0 & 0 & 2/3 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 & 0 \end{pmatrix}.$$

1. Draw the transition diagram for this Markov chain.
2. Suppose the chain starts in state 3 initially.
 - (i) What are the probabilities that state 1 is reached for the first time at the 1st, 2nd, and 3rd steps?
 - (ii) What is the probability that state 1 is ever reached (that is, what is h_{31})?
 - (iii) What is the probability that state 5 is ever reached (that is, what is h_{35})?
3. Suppose now that the chain starts in state 2 initially. What are the probabilities that states 5, 3 and 1 are ever reached?
4. Let $A = \{1, 2, 4, 5\}$. Find m_{3A} , the expected number of steps to reach the set of states A , given that the process starts in state 3.

11.3.7 In a production line, the colour of each item is independent and equally likely to be red (R), blue (B), green (G) or yellow (Y). Let Y_n = the colour of the n^{th} item ($= R, B, G$ or Y). A special sensor recognises when the sequence R, G, B occurs.

$$\text{Let } X_n = \begin{cases} 3 & \text{if } Y_{n-2} = R, Y_{n-1} = G, Y_n = B \\ 2 & \text{if } Y_{n-1} = R, Y_n = G \\ 1 & \text{if } Y_n = R \\ 0 & \text{otherwise.} \end{cases}$$

1. Obtain the transition matrix for the Markov chain $\{X_n\}$.
2. Find the equilibrium distribution π for this Markov chain.
3. Suppose the chain X is in state 3. What is the expected number of items that are produced until the chain is in state 3 again?

11.3.8 A spider and a fly move on the vertices of a square. At each time step they each move, independently of one another. With probability $1/3$ each one moves to the left, with probability $1/3$ it moves to the right, and with probability $1/3$ it stays where it is. Let the state of the system be the distance (in steps) between them, so that $S = \{0, 1, 2\}$.

- Two rows of the transition matrix are given below. Fill in the remaining row.

$$P = \begin{pmatrix} 1 & 0 & 0 \\ \frac{2}{9} & \frac{5}{9} & \frac{2}{9} \\ - & - & - \end{pmatrix}$$

- Assuming that the spider and the fly start at opposite corners of the square, what is the expected time until the spider meets (eats) the fly?

11.3.9 A lecturer distributes chocolate fish, of both the milk chocolate and dark chocolate variety.

- If she handed out a milk chocolate fish last time, she hands out a milk chocolate fish with probability $1/4$ this time. On the other hand, if she handed out a dark chocolate fish last time, she hands out a dark chocolate fish with probability $2/3$ this time.
 - What is the probability, in equilibrium, that a student receives a dark chocolate fish?
 - Suppose she has just handed out a dark chocolate fish. What is the expected number of fish that she hands out, before the next dark chocolate fish (include that dark chocolate fish in the count)?
 - In equilibrium, what is the probability that she hands out 2 dark chocolate fish in a row?
- Now suppose that she finds that she is running through her stock of dark chocolate fish too fast. She decides that she will follow the rule above, provided she has not handed out more than 3 chocolate fish in a row. If she ever hands out 3 chocolate fish in a row, the next fish will be milk chocolate with probability 1.
 - Explain why this system cannot be modelled by the same 2-state Markov chain as above.
 - Give an appropriate state space and transition diagram for this process.
 - What is the probability under the new rules that a student receives a dark chocolate fish, in equilibrium?
 - What is the probability under the new rules that she hands out 2 dark chocolate fish in a row?

Formulae

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

$$\text{Bayes' Formula:} \quad \mathbb{P}(B_j | A) = \frac{\mathbb{P}(A | B_j) \mathbb{P}(B_j)}{\sum_{i=1}^m \mathbb{P}(A | B_i) \mathbb{P}(B_i)},$$

if B_1, B_2, \dots, B_m form a partition of the sample space.

Discrete random variables

$$f_X(x) = \mathbb{P}(X = x), \quad F_X(x) = \mathbb{P}(X \leq x)$$

$$\mathbb{E}(g(X)) = \sum_x g(x) f_X(x), \quad \mathbb{E}(aX + b) = a\mathbb{E}(X) + b, \quad \mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$$

$$\text{Var}(X) = \mathbb{E}\left((X - \mathbb{E}(X))^2\right) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2, \quad \text{Var}(aX + b) = a^2 \text{Var}(X)$$

Joint distributions

$$f_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y), \quad \mathbb{E}(g(X, Y)) = \sum_{x,y} g(x, y) f_{X,Y}(x, y)$$

$$\text{Cov}(X, Y) = \mathbb{E}\left((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\right) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

$$\text{Cov}(aX + b, cY + d) = ac \text{Cov}(X, Y)$$

$$\text{Cov}(X + Y, Z + W) = \text{Cov}(X, Z) + \text{Cov}(X, W) + \text{Cov}(Y, Z) + \text{Cov}(Y, W)$$

Conditional distributions

$$f_{X|A}(x) = \mathbb{P}(X = x | A), \quad \mathbb{E}(g(X) | A) = \sum_x g(x) f_{X|A}(x)$$

$$f_{Y|X}(y|x) = \mathbb{P}(Y = y | X = x) = \frac{f_{X,Y}(x, y)}{f_X(x)}$$

$$\mathbb{E}(Y | X) = \psi_Y(X), \quad \text{where} \quad \psi_Y(x) = \mathbb{E}(Y | X = x) = \sum_y y f_{Y|X}(y|x)$$

$$\mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y | X)), \quad \mathbb{E}(g(X)Y) = \mathbb{E}(g(X)\mathbb{E}(Y | X))$$

$$\text{Cov}(X, Y) = \text{Cov}(X, \mathbb{E}(Y | X)), \quad \text{Var}(Y) = \mathbb{E}(\text{Var}(Y | X)) + \text{Var}(\mathbb{E}(Y | X))$$

Named distributions

Binomial:

$$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{(n-x)!x!} p^x (1-p)^{n-x}, \quad \text{for } x = 0, 1, \dots, n$$

$$\mathbb{E}(X) = np, \quad \text{Var}(X) = np(1-p)$$

Geometric:

$$f_X(x) = (1-p)^x p, \quad \text{for } x = 0, 1, 2, \dots$$

$$\mathbb{E}(X) = \frac{1-p}{p}, \quad \text{Var}(X) = \frac{1-p}{p^2}$$

Negative Binomial:

$$f_X(x) = \binom{k+x-1}{x} p^k (1-p)^x, \quad \text{for } x = 0, 1, 2, \dots$$

$$\mathbb{E}(X) = \frac{k(1-p)}{p}, \quad \text{Var}(X) = \frac{k(1-p)}{p^2}$$

Poisson:

$$f_X(x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad \text{for } x = 0, 1, 2, \dots$$

$$\mathbb{E}(X) = \text{Var}(X) = \lambda$$

Discrete Uniform:

$$f_X(x) = \frac{1}{N}, \quad \text{for } x = 1, 2, \dots, N$$

$$\mathbb{E}(X) = \frac{N+1}{2}, \quad \text{Var}(X) = \frac{N^2-1}{12}$$

Hypergeometric:

$$f_X(x) = \frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}}, \quad \text{for } x = \max(0, n-N+M), \dots, \min(n, M)$$

$$\mathbb{E}(X) = \frac{nM}{N}, \quad \text{Var}(X) = n \frac{M}{N} \left(1 - \frac{M}{N}\right) \frac{N-n}{N-1}$$

Other useful formulae

Summation

If c and d are constants,

$$\sum_i (ca_i + db_i) = c \sum_i a_i + d \sum_i b_i$$

Shifting the index of summation:

$$\sum_{i=1}^n a_{i+m} = \sum_{j=m+1}^{m+n} a_j$$

Reversing the index of summation:

$$\sum_{i=0}^n a_i b_{n-i} = \sum_{j=0}^n a_{n-j} b_j$$

Arithmetic series with first term a and common difference d :

$$T_i = a + (i-1)d, \quad S_n = \sum_{i=1}^n T_i = \frac{n}{2}(2a + (n-1)d)$$

In particular if $T_i = i$,

$$S_n = \sum_{i=1}^n i = \frac{n(n+1)}{2}$$

Geometric series with first term a and common ratio r :

$$T_k = ar^{k-1}, \quad S_n = \sum_{i=1}^n T_i = \frac{a(r^n - 1)}{r - 1} = \frac{a(1 - r^n)}{1 - r}$$

If $|r| < 1$ then

$$S_\infty = \sum_{i=1}^{\infty} T_i = \frac{a}{1 - r}$$

Factorials and binomial coefficients:

$$\begin{aligned} x! &= x(x-1)(x-2) \cdots 3 \times 2 \times 1 = x \times (x-1)!, & 1! &= 0! = 1 \\ {}^n P_r &= \frac{n!}{(n-r)!} = n(n-1)(n-2) \cdots (n-r+2)(n-r+1) \\ \binom{n}{r} &= {}^n C_r = \frac{{}^n P_r}{r!} = \frac{n!}{r!(n-r)!} \end{aligned}$$

Binomial theorem:

$$(x+y)^n = \sum_{k=0}^n \binom{n}{k} x^k y^{n-k}$$