Firstname Lastname

BSc Drinking Coffee (List Previous Degrees)

Title of the thesis should go here, line breaks may be necessary

Submitted 5 August, 1984
for the degree of Doctor of Philosophy

Music Computing and Psychology Lab,
Department of Music, University of York

Supervisors:

Firstname Lastname, Firstname Lastname,
and Firstname Lastname

Abstract

Summary of thesis to go here. One page maximum. Data and code related to this thesis is available from `https://osf.io/zv75h/`.

If typesetting this collection of LaTeX files locally (e.g., in TeXShop), you will need to typeset the main file thesis.tex once with the 'LaTeX' setting selected, then typeset it with the 'BibTeX' setting selected, then again with the 'LaTeX' setting selected. All paths to files are relative. The file Bibliography → bibliography.bib contains lots of examples of different kinds of references (e.g., PhD theses, books, papers in conference proceedings, a chapter in an edited book, etc.). You can remove a whole chapter at any point by commenting out an input command in thesis.tex (for example try turning line 135 into `% \input{"Evaluation/draft"}` and see what happens).

I have included a bunch of examples of mathematical typesetting in Appendix A (p. 17). The documentation called lshort.pdf by Tobias Oetiker is also an excellent guide.

My one major piece of advice is to click 'Recompile' (Overleaf) or 'Typeset' (e.g., TeXShop) regularly. In this way, you will catch errors just after you make them, and not have to spend too long finding and correcting them.

Acknowledgements

Thanks to go here.

Author Name,
City,
Month Year.

Related publications

If we get some peer-reviewed publications before you finish, you can list them. It is helpful to let your examiners know that certain content has already gone through a peer-review process. For example, Chapters 3 and 4 are based on the following journal paper.

> Firstname Lastname, Firstname Lastname, and Firstname Lastname. This is a publication title. *Journal Where Published*, 28(4):387-414, 2011.

Chapter 2 is based on the following conference paper.

> Firstname Lastname and Firstname Lastname. This is a publication title. In Firstname Lastname and Firstname Lastname, editors, *Proceedings of the Great Conference*, pages 3-8, City, Country (or City, State), Year. Publisher.

Contents

List of Tables

List of Figures

1 Introduction

General introduction, leading towards some research questions and hypotheses:

> Blah?

Perhaps this question can be broken down into two halves:

Question 1. Blah?

Question 2. Blah?

Interpret the questions and say how they will be addressed. In so doing, you will give an overview of the contents of the following chapters. Remember referencing systems, like Chapter 3, individual sections such as Sec. 4.2.2, and Figures (please see Fig. 1.1). You can also make references to authors in your bibliography, such as Collins (2011). Or in parentheses like this (Collins et al., 2010, 2013). You may prefer to use a numbered system instead [1, 2], in which case change the `\bibliographystyle` command from `plainnat` to `plain` (toward the bottom of thesis.tex).

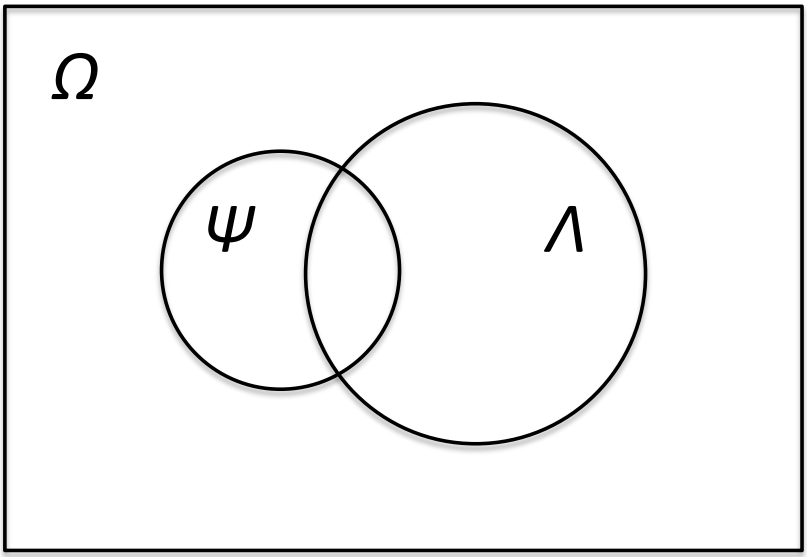

Figure 1.1: A longer caption can go here. If there are multiple parts to the figure, use something like this. (A) Caption specific to first labelled part; (B) Caption specific to second labelled part, and so on.

2 Literature review: Music representations

2.1 Musical instrument digital interface

Blah blah.

Please make sure all music examples are neatly typeset in MuseScore (`https:musescore.org`). When referring to a composer for the first time, use their full name and give their birth/death dates. Subsequent references can use the surname only, but include initials as well if these are necessary in order to distinguish one composer from another (for instance J.S. Bach and J.C. Bach).

Figure 2.1 gives an example of how to refer to an excerpt of a piece. Song titles should be in quotations, e.g., ‘Summertime’, and pieces with descriptive titles should be in italics, such as *The Rite of Spring*. Permission to use copyrighted material should be sought where necessary.

Graphs should be created with a homogeneous style in R, Matplotlib, or Nodeplotlib. The font size for text in the plot should appear as large as the font size in the main text.

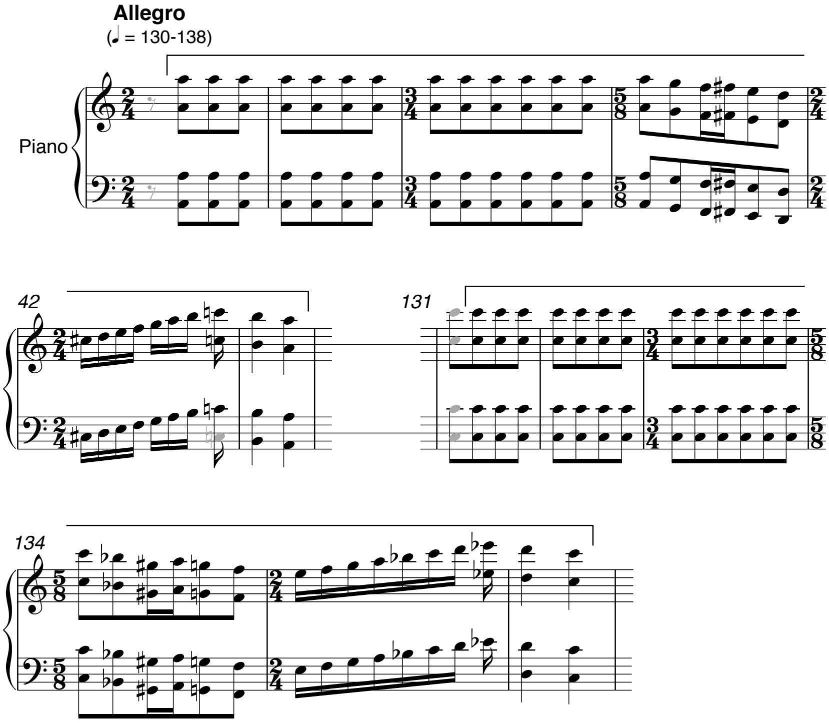

Figure 2.1: Bars 38-43 and 131-136 from the first movement of the Piano Concerto no.1 by Béla Bartók (1881-1945). The brackets indicate an instance of a real sequence. Black noteheads help to show which notes are involved. ⓒ Copyright 1927 by Universal Edition. Copyright renewed 1954 by Boosey & Hawkes, Inc., New York. Reproduced by permission.

2.2 The generalised interval system and viewpoints

Blah blah.

2.3 Geometric representations

Blah blah.

2.4 Some music-analytical tools

Blah blah.

3

Literature review:

Discovery of patterns in music

Blah blah.

4 Evaluation

4.1 Evaluation questions

This chapter consists of an evaluation of blah. The purpose of the evaluation is to address the following questions:

1. Blah?

2. Blah?

4.2 Methods for answering evaluation questions

Blah blah.

4.2.1 Participants

Blah blah.

4.2.2 Stimuli

Stimuli were prepared blah.

4.2.3 Procedure

Temporal account of what participants did.

4.2.4 Apparatus

Any special equipment used? Describe it here.

4.3 Results

4.3.1 Answer to evaluation question 1

Blah.

4.3.2 Answer to evaluation question 2

Blah blah.

4.4 Local conclusions

Table 4.1: Some data in a table and a longer caption. The data here are typeset in a slightly smaller font just as an example, but the longtable command is intended to allow for tables spreading over two or more pages.

Stim-	Multicolumn!		And again		Once more	
ulus	Yes	No	Random	Words	In	Titles
	Hello, some neat multicolumn action					
1	4.56	5.38	31.3	68.8	31.3	68.8
2	5.63	5.56	68.8	81.3	68.8	81.3
And again)						
9	3.06	2.00	81.3	87.5	12.5	6.3
11	1.19	1.38	68.8	81.3	0.0	0.0
12	3.06	2.69	31.3	68.8	12.5	0.0
Yet another category						
13	4.75	5.88	25.0	6.3	43.8	87.5
14	5.38	5.13	12.5	25.0	62.5	56.3
18	5.25	5.63	12.5	6.3	75.0	68.8
And so on						
20	4.75	4.38	25.0	62.5	$\boxed{56.3}$	$\boxed{25.0}$
21	2.81	2.69	75.0	81.3	6.3	0.0
24	3.13	3.19	68.8	93.8	18.8	0.0
And so on						
25	2.00	1.81	75.0	81.3	0.0	0.0
Finally, the last one!						
28	3.25	2.88	43.8	81.3	$\boxed{25.0}$	0.0
31	2.75	2.89	50.0	87.5	0.0	0.0
32	2.50	2.75	81.3	93.8	12.5	0.0

Table 4.2: Contrasts for two ANOVAs, a longer caption. The magnitude of the number indicates the significance of this difference in stylistic success. One, two, and three asterisks indicate significance at the .05, .01, and .001 levels respectively, testing a two-sided hypothesis using a $t(26)$ distribution.

Concertgoers

Source	System B	System C	System D	System E	System F
System A	0.429	0.844	0.844	2.500***	2.865***
System B	.	0.830	0.830	4.145***	4.875***
System C	.	.	0.000	3.477***	4.242***
System D	.	.	.	3.477***	4.242***
System E	.	.	.	.	0.765

$F(5,26) = 10.12, \quad p = 1.827 \times 10^{-5}, \quad s = 0.825$

Experts

Source	System B	System C	System D	System E	System F
System A	0.421	0.417	0.677	2.875***	3.083***
System B	.	−0.008	0.468	4.485***	4.865***
System C	.	.	0.499	4.712***	5.111***
System D	.	.	.	4.212***	4.612***
System E	.	.	.	.	0.399

$F(5,26) = 12.16, \quad p = 3.874 \times 10^{-6}, \quad s = 0.904$

5

Conclusions and future work

5.1 Conclusions

Blah blah.

5.2 Future work

Blah blah.

Appendices

A
Mathematical definitions

This is an appendix to demonstrate the style for typesetting definitions, equations, etc. Please use it to find examples of mathematical typesetting that you need in the thesis. Be sure to remove it afterwards: it is pretty much an exact copy of one of the appendices in my PhD thesis (Collins, 2011).

Definition A.1. Vector. A vector is a collection of numbers, separated by commas and enclosed by parentheses '(' and ')'. A vector may contain the same number more than once. It is standard to use lowercase bold letters to denote vectors. ■

Example A.2. Here are some examples of vectors:

$$\mathbf{a} = (1, 2, 3), \quad \mathbf{b} = (2, 1, 3), \quad \mathbf{c} = (c_1, c_2, \ldots, c_n). \tag{A.1}$$

The vector $\mathbf{c}$ demonstrates the general notation for a vector, that is one to which numerical values have not been assigned. The ellipsis '...' is useful for saving time and space. The vectors $\mathbf{a}$ and $\mathbf{b}$ from (A.1) are *not* considered to be equal: they contain the same numbers, but in different orders. In general, two vectors $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ are said to be equal if $m = n$ and $x_i = y_i$, where $i = 1, 2, \ldots, m$. ■

Definition A.3. Matrix, matrix operations, and array. Whereas a vector is a list of numbers, a *matrix* is a table of numbers, consisting of m rows and n columns. The entry in the ith row, jth column of a matrix $\mathbf{A}$ is denoted $(\mathbf{A})_{i,j}$ or $a_{i,j}$. So

$$\mathbf{A} = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{pmatrix}. \tag{A.2}$$

The sum of two $m \times n$ matrices $\mathbf{A}$ and $\mathbf{B}$ is defined by $(\mathbf{A}+\mathbf{B})_{i,j} = (\mathbf{A})_{i,j} + (\mathbf{B})_{i,j}$. Similarly for subtraction. For a constant $\lambda \in \mathbb{R}$, $\lambda\mathbf{A}$ is defined by $(\lambda\mathbf{A})_{i,j} = \lambda(\mathbf{A})_{i,j}$. The *diagonal* of an $m \times n$ matrix $\mathbf{A}$ is a list consisting of the elements $a_{i,i}$, where $1 \leq i \leq \min\{m, n\}$. The *upper triangle* of $\mathbf{A}$ is a list consisting of the elements $a_{i,j}$, where $1 \leq i \leq \min\{m, n\}$ and $i < j$. The rth *superdiagonal* of a $\mathbf{A}$ is a list consisting of the elements $a_{i,i+r}$, where $1 \leq i \leq n - r$.

The product of $\mathbf{A}$, an $m \times n$ matrix, and $\mathbf{B}$, an $n \times p$ matrix, is written $\mathbf{AB}$, an $m \times p$ matrix, and its ith row, jth column is given by

$$(\mathbf{AB})_{i,j} = \sum_{k=1}^{n} a_{i,k} b_{k,j}. \tag{A.3}$$

Other matrix operations include *transposition* and *inversion.* For an $m \times n$ matrix $\mathbf{A}$, the *transpose* is written $\mathbf{A}^T$, and its ith row, jth column is given by

$$(\mathbf{A}^T)_{i,j} = (\mathbf{A})_{j,i}. \tag{A.4}$$

The identity matrix $\mathbf{I}$ is an $m \times m$ matrix such that $(\mathbf{I})_{i,j} = 1$ for $i = j$, and $(\mathbf{I})_{i,j} = 0$ otherwise. For $\mathbf{A}$, an $m \times n$ matrix, under certain conditions (not specified here) there exists $\mathbf{B}$, an $n \times m$ matrix, such that $\mathbf{AB} = \mathbf{I}$. In which case, we say that $\mathbf{B}$ is the *matrix inverse* of $\mathbf{A}$, and use the notation $\mathbf{A}^{-1} = \mathbf{B}$.

A one-dimensional *array* is a vector; a two-dimensional *array* is a matrix. It is possible to extend the concept of an array to d dimensions, although such arrays are not easily displayed on paper, and the index notation becomes unwieldy. Let us consider the case $d = 3$. We can define $\mathbf{A}^{(k)}$ to be an $m \times n$ matrix with ith row, jth column denoted $a_{i,j,k}$, and imagine stacking p matrices $\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \ldots, \mathbf{A}^{(p)}$ *back to back* to form an $m \times n \times p$ *block* of numbers. If we denote the stacked matrices by $\mathbf{A}$, then $\mathbf{A}$ is a three-dimensional array.

In Chapter 3, we might use the notation

$$\mathbf{A} = \begin{pmatrix} \mathbf{a}_{1,1} & \mathbf{a}_{1,2} & \cdots & \mathbf{a}_{1,n} \\ \mathbf{a}_{2,1} & \mathbf{a}_{2,2} & \cdots & \mathbf{a}_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{a}_{m,1} & \mathbf{a}_{m,2} & \cdots & \mathbf{a}_{m,n} \end{pmatrix} \tag{A.5}$$

for a three-dimensional array. That is, the element $a_{i,j,k}$ of the array $\mathbf{A}$ can be thought of as the kth element of the vector $\mathbf{a}_{i,j}$. ■

Definition A.4. String. A string is a collection of alphabetic characters enclosed by quotation marks ‘ and ’. For musical purposes, other admissible characters in a string are the accidental symbols ‘♮’, ‘♯’, ‘♭’, ‘𝄪’, and ‘𝄫’, as well as the space symbol ‘ ’. Similar to vectors, a string may contain the same character more than once and it is standard to use lowercase bold letters to denote strings. ■

Example A.5. Here are some examples of strings:

$$\mathbf{s} = \text{‘Piano’}, \quad \mathbf{t} = \text{‘Violin I’}, \quad \mathbf{u} = \text{‘ATGCAACT’}, \quad \mathbf{v} = \text{‘G}\sharp\text{’}. \qquad \text{(A.6)}$$

The comments in Example A.2 about general notation, the use of ellipses, and equality apply also to strings. ■

Definition A.6. Concatenation. For two strings $\mathbf{s}$ = ‘$s_1 s_2 \cdots s_m$’ and $\mathbf{t}$ = ‘$t_1 t_2 \cdots t_n$’, the notation $\text{conc}(\mathbf{s}, \mathbf{t})$ is used to mean the concatenation of the two strings, that is $\text{conc}(\mathbf{s}, \mathbf{t})$ = ‘$s_1 s_2 \cdots s_m t_1 t_2 \cdots t_n$’. ■

Definition A.7. List and set. A list is a collection of elements. Admissible elements of a list are numbers, vectors, strings, sets (see below), and lists themselves. Like vectors, the elements of a list are separated by commas and enclosed by parentheses ‘(’ and ‘)’. For a list, the order of elements matters as far as equality is concerned. A list may contain the same element more than once. It is standard to use uppercase italic letters to denote lists, and lowercase italic letters to denote their elements.

A set is a collection of elements. Admissible elements of a set are numbers, vectors, strings, lists, and sets themselves. The elements of a set are separated by commas and enclosed by curly brackets ‘{’ and ‘}’. Unlike vectors, strings, and lists, a set is unordered as far as equality is concerned, and must not contain repeated elements. As with lists, it is standard to use uppercase italic letters to denote sets, and lowercase italic letters to denote their elements. The notation $a \in A$ is used to mean a is an element of the set A. A set A

is said to be a subset of a set B if for each $a \in A$, $a \in B$. Two sets A and B are said to be equal if A is a subset of B, and B is a subset of A. The notation $A \subset B$ is used to mean that A is a subset of B but not equal to it, and $A \subseteq B$ to mean A is a subset of B or equal to it. ■

Example A.8. Here is an example of a list:

$$L = (3, 4, \mathbf{a}, 5, 3, (2, \mathbf{b}), \text{`Viola'}), \tag{A.7}$$

and here are several examples of sets:

$$A = \{2, 1, 3\}, \quad B = \{4, 3, 2\}, \quad C = \{1, 3, 2\}, \quad D = \{\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_n\}. \tag{A.8}$$

So $A = C$. Eventually I will run out of letters to represent numbers and sets, in which case the Greek alphabet may also be employed, as well as some kind of indexing system, as with D in (A.8). Unless stated otherwise, definitions are refreshed with each new numbered equation. That is, **a** and **b** from (A.7) do not bear any relation to **a** and **b** from (A.1). In fact, each could be a vector or a string. ■

Definition A.9. Union, intersection, set difference, and Cartesian product. The union of two sets A and B, written $A \cup B$, is the set of all elements x such that $x \in A$ or $x \in B$. The previous sentence can be expressed as set notation:

$$A \cup B = \{x : x \in A \text{ or } x \in B\}. \tag{A.9}$$

The 'or' is inclusive, meaning it is acceptable for x to be in both A and B.

The intersection of two sets A and B, written $A \cap B$, is the set of all elements x such that $x \in A$ and $x \in B$. That is,

$$A \cap B = \{x : x \in A \text{ and } x \in B\}. \tag{A.10}$$

The set difference of two sets A and B, written $A \backslash B$, is the set of all elements x such that $x \in A$ and $x \notin B$, where '$\notin$' means 'not in'. That is,

$$A \backslash B = \{x : x \in A \text{ and } x \notin B\}. \tag{A.11}$$

The Cartesian product of two sets A and B, written $A \times B$, is the set of all lists (a, b) such that $a \in A$ and $b \in B$. That is,

$$A \times B = \{(a, b) : a \in A, b \in B\}. \tag{A.12}$$

Each of these definitions (union, intersection, and Cartesian product) extend naturally to n sets $A_1, A_2, \ldots, A_n$. For instance,

$$A_1 \times A_2 \times \cdots \times A_n = \{(a_1, a_2, \ldots, a_n) : a_1 \in A_1, a_2 \in A_2, \ldots, a_n \in A_n\}. \tag{A.13}$$

Sometimes, Cartesian products over the same set are abbreviated. For instance, $A \times A \times A = A^3$. ■

Example A.10. Taking the definitions of A and B from (A.8),

$$A \cup B = \{1, 2, 3, 4\}, \quad A \cap B = \{2, 3\}, \quad A \backslash B = \{1\}. \tag{A.14}$$

Again taking the definition of A from (A.8), and letting $B = \{\text{'Fl'}, \text{'Hn'}\}$,

$$A \times B = \{(1, \text{'Fl'}), (1, \text{'Hn'}), (2, \text{'Fl'}), (2, \text{'Hn'}), (3, \text{'Fl'}), (3, \text{'Hn'})\}\,. \quad \text{(A.15)}$$

■

Definition A.11. Function. A function, represented by an italic letter such as f or a non-italic short word such as max or cos, is a collection of rules that describe how elements of one set A called the *domain* are mapped to elements of another set B. A mathematical shorthand for the previous sentence is $f : A \to B$. The set denoted $f(A)$ and defined by $f(A) = \{f(a) : a \in A\}$ is called the *image* of the function. ■

Example A.12. With A and B defined as in (A.8), an example of a function is

$$f(a) = \begin{cases} 2, & \text{if } a = 1, \\ 3, & \text{if } a = 2, \\ 4, & \text{if } a = 3. \end{cases} \quad \text{(A.16)}$$

The mathematics '$f(a)$' is read 'f of a'. Convention stipulates that the *argument*, an element a of the domain A, is placed within parentheses or square brackets to the right of the function name, in this case f. The function states that $1 \in A$ maps to $2 \in B$, $2 \in A$ maps to $3 \in B$, and $3 \in A$ maps to $4 \in B$. Alternatively, one could write $f(1) = 2$, $f(2) = 3$, and $f(3) = 4$. It would be more concise (and therefore preferable) to define $f : A \to B$ by

$$f(a) = a + 1, \quad a \in A. \quad \text{(A.17)}$$

Such concise definitions of a function are not always possible. For instance, with A and B defined as in (A.8), let $g : A \to B$ be given by

$$g(a) = \begin{cases} 2, & \text{if } a = 1, \\ 3, & \text{if } a = 1, \\ 2, & \text{if } a = 2. \end{cases} \tag{A.18}$$

This function defies attempts at concision. ■

Definition A.13. Well defined, onto, one-to-one, bijective, and invertible. A function $f : A \to B$ is said to be well defined if the mapping of each element $a \in A$ to $b \in B$ is unambiguous. (For example, f in A.16 is well defined, whereas g in A.18 is not well defined, as it is unclear whether $1 \in A$ should map to $2 \in B$ or $3 \in B$.) If for each element $b \in B$ of a function $f : A \to B$, there exists (at least) one elment $a \in A$ such that $f(a) = b$, then f is said to be *onto.* Another property that a function $f : A \to B$ might exhibit is *one-to-oneness.* If for each element $a_1 \in A$, there is *no other* element $a_2 \in A$ such that $f(a_1) = f(a_2)$, then f is said to be *one-to-one.*

A function $f : A \to B$ that is both one-to-one and onto is called *bijective.* A function $f : A \to B$ is said to be *invertible* if there exists a function $f^{-1} : B \to A$ such that $f(a) = b$ if and only if $f^{-1}(b) = a$. It can be shown (but will not be shown here) that a function f is invertible if and only if it is bijective. ■

Example A.14. Here are some more examples of functions, exhibiting various combinations of one-to-one and onto properties.

$$f_1 : \mathbb{R} \to \mathbb{R}, \text{ by } f_1(x) = x^2, \tag{A.19}$$

$$f_2 : \mathbb{Z} \to \mathbb{Z}, \text{ by } f_2(m) = m^3, \tag{A.20}$$

$$f_3 : \mathbb{R}^2 \to \mathbb{R}, \text{ by } f_3\left[(x, y)\right] = x + y, \tag{A.21}$$

$$f_4 : \mathbb{R} \to \mathbb{R}, \text{ by } f_4(x) = x^3, \tag{A.22}$$

$$f_5 : \mathbb{R}^n \to \mathbb{R}, \text{ by } f_5[(x_1, x_2, \ldots, x_n)] = \tfrac{1}{n}(x_1 + x_2 + \cdots + x_n), \tag{A.23}$$

$$f_6 : \mathbb{R}^n \to \mathbb{R}, \text{ by } f_6[(x_1, x_2, \ldots, x_n)] = (x_1 \cdot x_2 \cdots x_n)^{(1/n)}, \tag{A.24}$$

$$f_7 : \mathbb{R} \to [-1, 1], \text{ by } f_7(t) = t - t^3/3! + t^5/5! - t^7/7! + \cdots\,, \tag{A.25}$$

$$f_8 : \mathbb{R} \to [-1, 1], \text{ by } f_8(t) = 1 - t^2/2! + t^4/4! - t^6/6! + \cdots\,. \tag{A.26}$$

The function f_1 is neither one-to-one nor onto. Both 1^2 and $(-1)^2$ equal 1, so f_1 is not one-to-one. There is no real number x such that $x^2 = -1$, so f_1 is not onto. Without further explanation, f_2 is one-to-one but not onto, f_3 is not one-to-one but is onto, and f_4 is both one-to-one and onto. None of the functions f_5, f_6, f_7, f_8 are one-to-one, but they are all onto. The function f_5 is the arithmetic mean, and f_6 is the geometric mean, where '$\cdot$' is a more accepted sign than '$\times$' for multiplying numbers. Writing out the functions (A.23)-(A.26) in full each time can be cumbersome, so a shorthand called *sigma notation* is used. For example, (A.23) can be re-written as

$$f_5(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{A.27}$$

which reads 'f_5 of the vector $\mathbf{x}$ equals 1 divided by n times the sum from i

equals 1 to i equals n of x_i'. The arithmetic mean of a vector $\mathbf{x}$ is sometimes denoted $\overline{\mathbf{x}}$. Similarly,

$$f_6(\mathbf{x}) = \left(\prod_{i=1}^{n} x_i\right)^{(1/n)}, \tag{A.28}$$

It is harder to cajole the functions f_7 and f_8 into sigma notation, but here they are:

$$\sin(t) = f_7(\mathbf{t}) = \sum_{i=1}^{\infty} \frac{(-1)^{i-1}}{(2i-1)!}\, t^{2i-1}, \tag{A.29}$$

$$\cos(t) = f_8(\mathbf{t}) = \sum_{i=0}^{\infty} \frac{(-1)^{i}}{(2i)!}\, t^{2i}. \tag{A.30}$$

These functions are shown with their special names, sin (short for sine) and cos (short for cosine) respectively. ■

Definition A.15. Combination of functions. For two functions $f : A \to B$ and $g : B \to C$, the *combination* $f \circ g : A \to C$ is defined by $g(f(a))$, where $a \in A$. For n functions $f_1 : A_0 \to A_1, f_2 : A_1 \to A_2, \ldots, f_n : A_{n-1} \to A_n$, the *combination* $f_1 \circ f_2 \circ \cdots \circ f_n : A_0 \to A_n$ is defined by $f_n\left(f_{n-1}\left(\cdots\left(f_2\left(f_1\left(a_0\right)\right)\right)\cdots\right)\right)$, where $a_0 \in A_0$. Often this is called composition of functions, but the term *combination* will be used here, to avoid confusion with musical composition. ■

Example A.16. Let $f_1 : \mathbb{R}_+ \to \mathbb{R}_+$ be defined by $f_1(a_0) = 2\pi 440 a_0$, let $f_2 : \mathbb{R}_+ \to [-1, 1]$ be defined by $f_2(a_1) = \sin(a_1)$, and let $f_3 : [-1, 1] \to [-0.7, 0.7]$

be defined by $f_3(a_2) = 0.7a_2$. Then

$$f_1 \circ f_2 \circ f_3(a_0) = f_3\left(f_2\left(f_1\left(a_0\right)\right)\right) \tag{A.31}$$

$$= 0.7\sin\left(2\pi 440t\right) \tag{A.32}$$

$$= 0.7\left(\sum_{i=1}^{\infty} \frac{(-1)^{i-1}}{(2i-1)!}\left(2\pi 440a_0\right)^{2i-1}\right). \tag{A.33}$$

■

Definition A.17. Binary operator. A binary operator is a function $f : A^2 \to A$. It is common to see elements of the argument for a binary operator written either side of the function symbol, rather than to the right. That is, $x + y$ is equivalent to and more common than $f_3[(x, y)]$, where f_3 was defined in (A.21). The general symbol for a binary operator is '$\circ$', so one might see $x \circ y$. This should not be confused with the same symbol used for *combinations of functions* (Def. A.15). Sometimes the symbol is dropped altogether, so $xy = x \circ y$. Apart from addition over the real numbers, other examples of binary operators include subtraction and multiplication. ■

Definition A.18. Modulo arithmetic. It can be shown (but will not be shown here) that for $a \in \mathbb{N}$, an arbitrary integer $n \in \mathbb{Z}$ can be expressed uniquely as $n = am + b$, where $b, m \in \mathbb{Z}$, and $0 \leq b < a$. For example, fixing $a = 12$, we have $61 = 12 \cdot 5 + 1$, and $-7 = 12 \cdot (-1) + 5$. This fact is used to define a function $f : (\mathbb{Z} \times \mathbb{N}) \to \mathbb{Z}_a$ by $f\left[(n, a)\right] = b$, where $n = am + b$ for integers b, m, and $0 \leq b < a$. In words, it is said that 'n equals b modulo a'.

For two elements $x, y \in \mathbb{Z}_a$, the binary operator of addition modulo a, written '$+_a$', is defined by

$$x +_a y = \begin{cases} x + y, & \text{if } x + y < a, \\ x + y - a, & \text{otherwise.} \end{cases} \tag{A.34}$$

■

Definition A.19. Group. A *group* $(G, \circ)$ consists of a set G and a binary operation $\circ$, such that:

1. *Closure.* For all $x, y \in G$, $x \circ y \in G$.
2. *Associativity.* For all $x, y, z \in G$, $(x \circ y) \circ z = x \circ (y \circ z)$.
3. *Identity.* There exists $e \in G$ such that $e \circ x = x \circ e = x$, for all $x \in G$.
4. *Inverses.* For each $x \in G$, there exists an element written x^{-1} such that $x^{-1} \circ x = x \circ x^{-1} = e$. ■

Example A.20. It can be verified that each of $(\mathbb{R}, +)$, $(\mathbb{R}^*, \times)$, $(\mathbb{R}^*_+, \times)$, $(\mathbb{Q}, +)$, $(\mathbb{Q}^*, \times)$, $(\mathbb{Z}, +)$, and $(\mathbb{Z}_a, +_a)$ satisfy the conditions for closure, associativity, identity, and inverses given above, and so are groups.

Let x be defined as the clockwise rotation of a triangle about a point by 120°, let y be the same but by 240°, let e be the identity rotation (by 0°), and let the binary operator $\circ$ be defined as combinations of rotations, so that, for example, $x \circ x = x^2 = y$. Then letting $G = \{e, x, y\}$, it can be verified that $(G, \circ)$ is a group.

Another group $(G, \circ)$ consists of rotations of the cube that map vertices to vertices. Again, the binary operator is defined as combinations of rotations. The set G consists of twenty-four elements, one of which z is illustrated in

Fig. A.1. The left-hand side of Fig. A.1 shows a cube with vertices labelled $\omega_1, \omega_2, \ldots, \omega_8$. In the middle of Fig. A.1, an axis is drawn through vertices ω_1 and ω_7. If the cube is rotated by 120° about this axis as indicated by the arrow, then the vertices assume new positions, shown on the right-hand side of Fig. A.1. The next definition is motivated by the way in which the vertices of the cube are affected by such rotations. ■

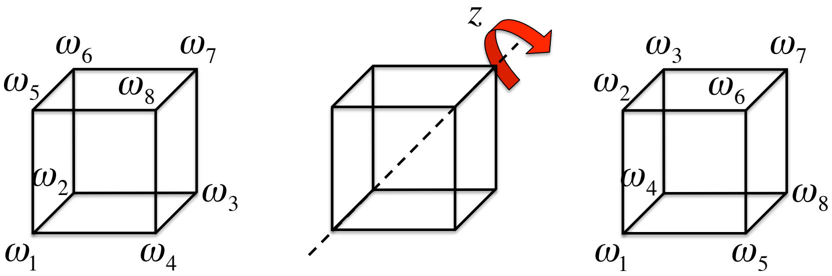

Figure A.1: The cube to the left has vertices labelled $\omega_1, \omega_2, \ldots, \omega_8$. The cube in the middle is subject to a rotation by 120° about the axis through ω_1 and ω_7. The cube to the right shows the vertices in their post-rotation positions.

Definition A.21. Action of a group on a set. Let $(G, \circ)$ be a group and Ω be a set. We say that G acts on Ω if the function $f : G \times \Omega \to \Omega$ satisfies the following conditions for each $\omega \in \Omega$:

1. For the identity element $e \in G$, $f(e, \omega) = \omega$.

2. For all $x, y \in G$, $f\left(x, f(y, \omega)\right) = f(x \circ y, \omega)$. ■

Example A.22. If, as above, the group $(G, \circ)$ consists of rotations of the cube that map vertices to vertices, and $\Omega = \{\omega_1, \omega_2, \ldots, \omega_8\}$ is the set of cube vertices, then G acts on Ω. With $z \in G$ defined as the rotation by 120° as illustrated in Fig. A.1, we have $f(z, \omega_1) = \omega_1$, and $f(z, \omega_2) = \omega_5$, and so on.

If there is a bijection $b : \Omega \to G$ for a set Ω and a group $(G, \circ)$, then G acts on Ω via the function $f : G \times \Omega \to \Omega$, defined by $f(x, \omega) = b^{-1}(x \circ b(\omega))$. ■

Definition A.23. Equivalence relation. A *relation* on a set S is a subset R of $S \times S$, indicating the ordered pairs of elements of S that are *related.* For $(s, t) \in R$, we write $s \sim t$, meaning s and t are related.

A relation is said to be:

- *Reflexive* if $s \sim s$ for all $s \in S$.
- *Symmetric* if $s \sim t$ implies $t \sim s$ for all $s, t \in S$.
- *Transitive* if ($s \sim t$ and $t \sim u$) implies $s \sim u$ for all $s, t, u \in S$.

An *equivalence relation* on a set S is a relation that is reflexive, symmetric, and transitive. For an equivalence relation R on a set S, two elements $s, t \in S$ such that $s \sim t$ are said to be in the same *equivalence class.* ■

Example A.24. Let $S = \mathbb{Z}$, $s, t \in S$, and R be a relation on S such that $s \sim t$ if $s \leq t$. It can be checked that the relation *is* reflexive, is *not* symmetric, and *is* transitive.

Now let $S = \mathbb{R}^2$ be the set of points in the plane, $(s_x, s_y), (t_x, t_y) \in S$, and R be a relation on S such that $(s_x, s_y) \sim (t_x, t_y)$ if $\sqrt{s_x^2 + s_y^2} = \sqrt{t_x^2 + t_y^2}$. In words, the point (s_x, s_y) is related to the point (t_x, t_y) if they are the same distance from the origin. It can be checked that this *is* an equivalence relation, and each *equivalence class* is a circle with centre the origin. ■

Definition A.25. Sample correlation coefficient. The sample correlation coefficient (also known as the Pearson product-moment correlation coefficient) of two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ is a function $f : (\mathbb{R}^n \times \mathbb{R}^n) \to [-1, 1]$ given by

$$f[(\mathbf{x}, \mathbf{y})] = \frac{\sum_{i=1}^{n}(x_i - \overline{\mathbf{x}})(y_i - \overline{\mathbf{y}})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{\mathbf{x}})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{\mathbf{y}})^2}}. \tag{A.35}$$

■

Example A.26. For the vectors $\mathbf{x} = (-8, 7, 2, 0)$ and $\mathbf{y} = (-5, 4, 0, 1)$, the sample correlation coefficient is $f[(\mathbf{x}, \mathbf{y})] = 0.971$. Keeping $\mathbf{x}$ the same and letting $\mathbf{z} = (9, -6, 4, 4)$, the sample correlation coefficient is $f[(\mathbf{x}, \mathbf{z})] = -0.923$. Keeping $\mathbf{x}$ the same and letting $\mathbf{w} = (1, 0, 8, -4)$, the sample correlation coefficient is $f[(\mathbf{x}, \mathbf{w})] = 0.072$.

So the sample correlation coefficient measures the strength of the linear relationship between two vectors, returning values close to 1 for a positive linear relationship, values close to -1 for a negative linear relationship, and values close to 0 for no linear relationship. ■

Definition A.27. Countable and cardinality. A set A is said to be *countable* (or *countably infinite*) if there exists a one-to-one function $f : A \to \mathbb{N}$. Otherwise it is *uncountable.* A set $A = \{a_1, a_2, \ldots, a_n\}$ with a finite number of elements is said to have *cardinality* $n = |A|$. ■

Example A.28. The sets $A = \{2, 1, 3\}$ and $B = \{b_1, b_2, \ldots, b_n\}$ are countable. The sets $\mathbb{Z}$ and $\mathbb{Q}$ are countable. In the latter case, there is an elegant proof that consists of constructing the matrix

$$\mathbf{A} = \begin{pmatrix} \frac{1}{1} & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \cdots \\ \frac{2}{1} & \frac{2}{2} & \frac{2}{3} & \frac{2}{4} & \frac{2}{5} & \cdots \\ \frac{3}{1} & \frac{3}{2} & \frac{3}{3} & \frac{3}{4} & \frac{3}{5} & \cdots \\ \frac{4}{1} & \frac{4}{2} & \frac{4}{3} & \frac{4}{4} & \frac{4}{5} & \cdots \\ \frac{5}{1} & \frac{5}{2} & \frac{5}{3} & \frac{5}{4} & \frac{5}{5} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{A.36}$$

The list $(\frac{1}{1}, \frac{2}{1}, \frac{1}{2}, \frac{1}{3}, \frac{2}{2}, \frac{3}{1}, \frac{4}{1}, \frac{3}{2}, \ldots)$ is formed by tracing a line over successive diagonals of $\mathbf{A}$. If each element $\frac{a}{b}$ in the list is proceeded by $-\frac{a}{b}$, and zero placed at the very beginning, then the rational numbers $\mathbb{Q}$ have been put in a one-to-one correspondence with the natural numbers $\mathbb{N}$.

The irrational numbers $\mathbb{I}$ and real numbers $\mathbb{R}$ are uncountable. Intervals, such as (a, b) and $[a, b]$, are uncountable. ■

Definition A.29. Sample space and event (Ross, 2006). The sample space of an experiment, denoted S, is the set of all possible outcomes. An event E is a subset of the sample space. The event E is said to have occurred if the experiment's outcome is contained in E. ■

Example A.30. Ross (2006). In an experiment that consists of rolling two dice, the sample space consists of thirty-six vectors

$$S = \big\{(i, j) : i, j \in \{1, 2, \ldots, 6\}\big\}, \tag{A.37}$$

where the outcome (i, j) occurs if i appears on the leftmost die and j on the rightmost die.

In an experiment that consists of measuring the lifetime of a transistor in hours, the sample space consists of all nonnegative real numbers

$$S = \{x \in \mathbb{R} : x \geq 0\}. \tag{A.38}$$

■

Definition A.31. Union, intersection, complement, and mutual exclusivity (Ross, 2006). The union and intersection of two sets were defined in Def. A.9. These definitions apply also to events, and can be extended from two to a countable number of events using a form of sigma notation (cf. p. 25) as follows. If $E_1, E_2, \ldots$ are events, the union of these events, denoted by $\bigcup_{i=1}^{\infty} E_i$, is defined to be that event consisting of all outcomes that are in E_i for at least one value of i, where $i = 1, 2, \ldots$. Similarly, the intersection of these events, denoted by $\bigcap_{i=1}^{\infty} E_i$, is defined to be the event consisting of those outcomes that are in all of the events E_i, where $i = 1, 2, \ldots$.

For an event E, the event E^{C}, called the *complement* of E, contains all events in the sample space S that are not in E. Recalling the definition of *set difference* (Def. A.9), $E^{\mathsf{C}} = S \backslash E$.

For two events E and F, if $E \cap F = \emptyset$, where $\emptyset$ is the empty set, then E and F are said to be *mutually exclusive.* ■

Axioms A.32. Axioms of probability (Ross, 2006).

1. For an experiment with sample space S, and an arbitrary event $E \subseteq S$, there exists a well defined function $\mathbb{P} : E \to [0, 1]$.

2. $\mathbb{P}(S) = 1$.

3. For arbitrary, mutually exclusive events $E_1, E_2, \ldots$, that is $E_i \cap E_j = \emptyset$ when $i \neq j$,
$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(E_i). \tag{A.39}$$

The notation $\mathbb{P}(E)$ is referred to as the probability of the event E. Results such as $\mathbb{P}(E^{\complement}) = 1 - \mathbb{P}(E)$ can be derived from the axioms. ■

Example A.33. If two fair dice are rolled, what is the probability that the sum of the upturned faces equals 8?

Solution. The state space for rolling two dice was given in Example A.30, and the events of interest are $E_1 = (2,6)$, $E_2 = (3,5)$, $E_3 = (4,4)$, $E_4 = (5,3)$, and $E_5 = (6,2)$. The five events are mutually exclusive and equiprobable, each with probability $\frac{1}{36}$. So the desired probability is $\frac{5}{36}$. ■

Definition A.34. Conditional probability and Bayes' formula (Ross, 2006). For two events E and F, if $\mathbb{P}(F) > 0$, then

$$\mathbb{P}(E \mid F) = \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(F)}. \tag{A.40}$$

The left-hand side of this equation reads 'the probability that the event E occurs given (or conditional on) the event F having occurred'.

Now suppose that $F_1, F_2, \ldots, F_n$ are mutually exclusive events such that $\bigcup_{i=1}^{n} F_i = S$, where S is the sample space. Then for some event E, it can be shown (but will not be shown here) that

$$\mathbb{P}(E) = \sum_{i=1}^{n} \mathbb{P}(E \mid F_i)\mathbb{P}(F_i). \tag{A.41}$$

Equation (A.41) is sometimes called the *law of total probability.*

Bayes' formula. With E and $F_1, F_2, \ldots, F_n$ defined as above,

$$\mathbb{P}(F_j \mid E) = \frac{\mathbb{P}(E \cap F_j)}{\mathbb{P}(E)} \tag{A.42}$$

$$= \frac{\mathbb{P}(E \mid F_j)\mathbb{P}(F_j)}{\sum_{i=1}^{n} \mathbb{P}(E \mid F_i)\mathbb{P}(F_i)}, \tag{A.43}$$

where $j \in \{1, 2, \ldots, n\}$ is arbitrary. Equation (A.43), Bayes' formula, follows from (A.40) and (A.41). ■

Example A.35. Adapted from Ross (2006). A note is played and a listener is asked to declare the pitch of the note. The listener is able to try out (play) *one* pitch before answering, and is told of three equally likely possibilities. We assume the listener is competent to the extent that if they try out an incorrect pitch, they will not declare that pitch. Let $1 - \beta_i$ denote the probability that the listener tries out and declares the ith pitch to be that of the note, when in fact this is correct, $i = 1, 2, 3$. The quantities $\beta_1, \beta_2, \beta_3$ are sometimes referred to as *overlook probabilities*. What is the conditional probability that the ith pitch is that of the note, given the listener tries out but does not declare the first pitch to be that of the note?

Solution. Let F_i, $i = 1, 2, 3$, be the event that the ith pitch is that of the note, and let E be the event that the listener tries out but does not declare the first pitch to be that of the note. From Bayes' formula (A.43),

$$\mathbb{P}(F_1 \mid E) = \frac{\mathbb{P}(E \cap F_1)}{\mathbb{P}(E)} \tag{A.44}$$

$$= \frac{\mathbb{P}(E \mid F_1)\mathbb{P}(F_1)}{\sum_{i=1}^{3} \mathbb{P}(E \mid F_i)\mathbb{P}(F_i)} \tag{A.45}$$

$$= \frac{\beta_1 \cdot \frac{1}{3}}{\beta_1 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3}} \tag{A.46}$$

$$= \frac{\beta_1}{\beta_1 + 2}. \tag{A.47}$$

For $j = 2, 3$,

$$\mathbb{P}(F_j \mid E) = \frac{\mathbb{P}(E \cap F_1)}{\mathbb{P}(E)} \tag{A.48}$$

$$= \frac{1 \cdot \frac{1}{3}}{\beta_1 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3}} \tag{A.49}$$

$$= \frac{1}{\beta_1 + 2}. \tag{A.50}$$

It is worth pointing out that the amount in (A.47) is less than one third, and the amount in (A.50) is more than one third. This makes intuitive sense: if the listener tries out but does not declare the first pitch, then the initial probabilities of the ith pitch being correct $(= \frac{1}{3},\ i = 1, 2, 3)$ are updated in favour of the second and third pitches. Also, the closer the overlook probability β_1 is to one, the closer the amounts in (A.47) and (A.50) are to one third. ■

Definition A.36. Independent events (Ross, 2006). Two events E and F are said to be *independent* if $\mathbb{P}(E \cap F) = \mathbb{P}(E) \cdot \mathbb{P}(F)$. Two events E and F that are not independent are said to be *dependent*.

The events $E_1, E_2, \ldots, E_n$ are said to be independent if for any subset $\{E_{i_1}, E_{i_2}, \ldots, E_{i_m}\}$ of them,

$$\mathbb{P}\left(\bigcap_{j=1}^{m} E_{i_j}\right) = \mathbb{P}(E_{i_1}) \cdot \mathbb{P}(E_{i_2}) \cdots \mathbb{P}(E_{i_m}). \tag{A.51}$$

■

Example A.37. Ross (2006). Suppose, as in Example A.30, that two fair dice are rolled. Let E be the event that the sum of the dice is 8, and F be the event that the leftmost dice shows 3. Then

$$\mathbb{P}(E \cap F) = \mathbb{P}\big(\{(3,5)\}\big) = \tfrac{1}{36}. \tag{A.52}$$

Having determined $\mathbb{P}(E) = \frac{5}{36}$ in Example A.33,

$$\mathbb{P}(E) \cdot \mathbb{P}(F) = \tfrac{5}{36} \cdot \tfrac{1}{6} = \tfrac{5}{216}. \tag{A.53}$$

Therefore, E and F are dependent. Suppose we let E be the event that the sum of the dice is 7. Now

$$\mathbb{P}(E \cap F) = \mathbb{P}\big(\{(3,4)\}\big) = \tfrac{1}{36}, \tag{A.54}$$

and

$$\mathbb{P}(E) \cdot \mathbb{P}(F) = \tfrac{1}{6} \cdot \tfrac{1}{6} = \tfrac{1}{36}. \tag{A.55}$$

Therefore, E and F are independent. ■

Definition A.38. Discrete random variable and probability mass function (Ross, 2006). Let S be a sample space and $E_1, E_2, \ldots$ be mutually exclusive events such that $\bigcup_{i=1}^{\infty} E_i = S$. A *discrete random variable* is a function $X : E_i \to \mathbb{R}$, well-defined for each value of $i = 1, 2, \ldots$. An arbitrary element of the image of X is denoted by a lowercase letter, such as x, or $x_1, x_2, \ldots$ if there are many. When an event E_i from the sample space is observed as the outcome, it is said that the random variable X takes or assumes a value x. The probability of the event E_i, denoted $\mathbb{P}(E_i)$, is equal to the probability that X takes or assumes the value x, written $\mathbb{P}(X = x)$.

The *probability mass function* of a discrete random variable X is defined by

$$p(x) = \mathbb{P}(X = x). \tag{A.56}$$

The domain of the function p is the countable set of values $\{x_1, x_2, \ldots\}$ that X can take. A probability mass function inherits properties from the Axioms of Probability (cf. Def. A.32). First, $p(x_i) \geq 0$, where $i = 1, 2, \ldots$. Second, $\sum_{i=1}^{\infty} p(x_i) = 1$. ■

Example A.39. Ross (2006). Often we are interested in some function of the outcome of an experiment, rather than the actual outcome itself. For instance, when rolling two dice, we might be interested in the sum of the two dice, and not really concerned about the separate values of each die. Random variables enable focus on a function of the experiment's outcome. Letting X be a random variable for the sum of two rolled dice, we have

$$\mathbb{P}(X = 2) = \mathbb{P}\big(\{(1,1)\}\big) = \tfrac{1}{36}, \tag{A.57}$$

$$\mathbb{P}(X = 3) = \mathbb{P}\big(\{(1,2),(2,1)\}\big) = \tfrac{1}{18}, \tag{A.58}$$

$$\mathbb{P}(X = 4) = \mathbb{P}\big(\{(1,3),(2,2),(3,1)\}\big) = \tfrac{1}{12}, \tag{A.59}$$

and so on. ■

Definition A.40. Bernoulli and binomial random variables (Ross, 2006). Suppose that the outcome of an experiment is either a success, in which case the discrete random variable X takes the value 1, or a failure, in which case X takes the value 0. Then the probability mass function of X is

$$p(x) = \mathbb{P}(X = x) = \begin{cases} 1 - \theta, & \text{if } x = 0, \\ \theta, & \text{if } x = 1, \end{cases} \tag{A.60}$$

where $0 \leq \theta \leq 1$ is the probability that the outcome of the experiment is a success. The random variable X is called a Bernoulli random variable.

Now suppose that in n independent experiments, each experiment has a successful outcome with probability θ, and failed outcome with probability $1 - \theta$. A discrete random variable Y that represents the number of successes that occur in n experiments is called a binomial random variable with parameters n, θ. We write $Y \sim B(n, \theta)$ as a shorthand to mean that Y is binomially distributed with parameters n, θ. The probability mass function is

$$p(i) = \binom{n}{i} \theta^i (1 - \theta)^{n-i}, \quad i = 0, 1, \ldots, n. \tag{A.61}$$

■

Definition A.41. Expectation, variance, and entropy of a discrete random variable. Let X be a discrete random variable taking the values $x_1, x_2, \ldots$, and let X have the probability mass function p. Then the *expectation* (also called the mean) of X, denoted $\mathbb{E}(X)$, is

$$\mu = \mathbb{E}(X) = \sum_{i=1}^{\infty} x_i p(x_i). \tag{A.62}$$

The expectation is a weighted average of the values assumed by X. The *variance* of X, denoted $\mathbb{V}(X)$, is

$$\mathbb{V}(X) = \mathbb{E}[(X-\mu)^2] = \mathbb{E}(X^2) - \mu^2. \tag{A.63}$$

The variance quantifies the average square distance between X and its mean.

Sometimes, we talk about a probability mass function as a *probability vector*. That is, the vector $\mathbf{p}$ with ith element $p_i = p(x_i)$, where $i = 1, 2, \ldots$. For the discrete random variable X with probability vector $\mathbf{p} = (p_1, p_2, \ldots, p_n)$, the *entropy* of X, denoted $H(X)$, is

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i, \tag{A.64}$$

where $-\log_2 p_i$ is known as the *information content* (Shannon, 1948). The entropy of a random variable quantifies the *uncertainty* associated with its outcome, with small positive values for low uncertainty, and large positive values for high uncertainty. ■

Example A.42. Suppose that X is a discrete random variable representing the sum of two rolled, *fair* dice. In Example A.39 we began calculating the probability mass function of X, which will now be given as a probability vector,

$$\mathbf{p} = \left(\frac{1}{36}, \frac{1}{18}, \frac{1}{12}, \frac{1}{9}, \frac{5}{36}, \frac{1}{6}, \frac{5}{36}, \frac{1}{9}, \frac{1}{12}, \frac{1}{18}, \frac{1}{36}\right). \tag{A.65}$$

After some calculations, we find $\mathbb{E}(X) = 7$, $\mathbb{V}(X) \approx 5.83$, and $H(X) \approx 3.27$.

Now suppose that Y is a discrete random variable representing the sum of two rolled dice, where one die is fair and the other biased towards higher scores, so that it shows $1, 2, \ldots, 6$ with respective probabilities $\frac{1}{32}, \frac{1}{32}, \frac{1}{16}, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}$. The probability vector $\mathbf{q}$ for Y is

$$\mathbf{q} = \left(\frac{1}{192}, \frac{1}{96}, \frac{1}{48}, \frac{1}{24}, \frac{1}{12}, \frac{1}{6}, \frac{31}{192}, \frac{5}{32}, \frac{7}{48}, \frac{1}{8}, \frac{1}{12}\right). \tag{A.66}$$

After some calculations, we find $\mathbb{E}(Y) \approx 8.53$, $\mathbb{V}(Y) \approx 4.57$, and $H(Y) \approx 3.07$. The biased die causes the expected value of Y to increase slightly compared with that of X. At the same time, the variance and entropy of Y are smaller respectively than the variance and entropy of X.

In general, it is possible to redistribute the mass of a probability vector $\mathbf{p}$, giving $\mathbf{q}$, such that for the corresponding random variables X and Y, $\mathbb{V}(X) < \mathbb{V}(Y)$, and $H(X) > H(Y)$. ■

Example A.30 contains an experiment where the lifetime of a transistor is measured in hours. The sample space was all nonnegative real numbers. The nonnegative real numbers, $\mathbb{R}_+$, are *uncountable* (cf. Def. A.27), just like the real numbers, $\mathbb{R}$. Discrete random variables cannot be used to model the exact outcome of such experiments, as their image must be a *countable* set. Another type of random variable is required, called a *continuous random variable.*

Definition A.43. Continuous random variable and probability density function (Ross, 2006). We say that X is a *continuous random variable* if there exists a nonnegative function f, defined for all $x \in \mathbb{R}$, such that for each set $A \subseteq \mathbb{R}$,

$$\mathbb{P}(X \in A) = \int_A f(x)\,\mathrm{d}x, \tag{A.67}$$

that is, the area between the curve $f(x)$ and the x-axis over which A is defined. The function f is known as the *probability density function* of X. It has the property that

$$1 = \mathbb{P}(X \in \mathbb{R}) = \int_{-\infty}^{\infty} f(x)\,\mathrm{d}x, \tag{A.68}$$

as X must belong to some interval.

Probability statements concerning X are answered in terms of f. For example, the probability that X takes a value in the interval $[a, b]$ is given by

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)\,\mathrm{d}x. \tag{A.69}$$

Definitions for the expectation, variance, and entropy of a continuous random variable are analogous to the discrete definitions, replacing sums with integrals. ■

Example A.44. The lifetime of a transistor measured in hours can be modelled by a continuous random variable X with probability density function

$$f(x) = \begin{cases} \frac{1}{\mu} e^{-x/\mu}, & \text{if } x \geq 0, \\ 0, & \text{if } x < 0, \end{cases} \tag{A.70}$$

where $\mu > 0$ is an arbitrary constant. Supposing a value of $\mu = 100$, what is the probability that a transistor will work between 80 and 140 hours before breaking?

Solution.

$$\mathbb{P}(80 \leq X \leq 140) = \int_{80}^{140} \frac{1}{100} e^{-x/100} \, \mathrm{d}x \tag{A.71}$$

$$= -e^{-x/100}\Big|_{80}^{140} \tag{A.72}$$

$$= e^{-0.8} - e^{-1.4} \tag{A.73}$$

$$\approx .203. \tag{A.74}$$

■

Definition A.45. Normal random variable (Ross, 2006). We say that X is a normal random variable, or that X is normally distributed, with parameters μ and σ^2 if the density of X is given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}, \tag{A.75}$$

where $x \in \mathbb{R}$. The notational shorthand $X \sim N(\mu, \sigma^2)$ means that X is normally distributed with parameters μ and σ^2. ■

Questions of a statistical nature are often answered in terms of the normal distribution, or in terms of related distributions, such as the t-distribution or F-distribution. There is an established method called *hypothesis testing* for stating and addressing questions concerning statistical significance. The following example gives a flavour for hypothesis tests, with more details being available elsewhere (Lunn, 2007a; Daly et al., 1995).

Example A.46. A previously unknown collection of Baroque 'cello concertos claimed to be by Antonio Vivaldi (1678-1741) is bequeathed to a library. The

Table A.1: The rhythmic density of various opening movements from known and supposed Vivaldi 'cello concertos. Data fabricated for the purpose of the example.

Bequeathed concertos		Library concertos		
6.44	1.86	5.14	3.15	6.15
3.66	3.67	4.19	4.29	5.29
4.58	4.04	5.37	4.46	5.81
5.06	3.32	4.75	5.33	4.42
3.09		4.69	4.82	
5.14		6.76	5.10	

librarian wishes to test this claim, so for each bequeathed concerto and for each Vivaldi 'cello concerto already held by the library, they calculate the rhythmic density of the opening movement. We can assume that rhythmic density is an appropriate aspect of the music to quantify, and fabricate some data for the purpose of the example (see Table A.1). The librarian wants to know whether these two sets of measurements constitute evidence of a different composer. The so-called *null hypothesis* (sometimes denoted H_0) is that the two samples have underlying distributions with the same mean. The *alternative hypothesis* (or H_1) is that the two samples have underlying distributions with different means.

The difference between the means of each sample is $-0.897 = 4.086 - 4.983$. Is this difference significant, taking into consideration the size of and variation within each sample? We will not go into the details, but the difference in means is weighted by 0.416, and the ratio $-0.897/0.416 \approx -2.155$ is supposed to be an observation from a t-distribution (denoted T and similar to the normal distribution in Def. A.45) with twenty-four so-called degrees of freedom. It can be checked that $p = \mathbb{P}(|T| > 2.155) = .041$, so there is only a probability of .041 that the two samples have underlying distributions with the same mean. A typical cutoff point for rejecting the null hypothesis in favour of the alternative is $\alpha = .05$. As we have observed a p-value less than α, the null hypothesis is rejected. Giving a musical interpretation of this statistical result, there is evidence that the two sets of concertos may be by different composers. ■

Example A.47. Multiple linear regression. We may have cause to consider *linear models* of the form

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p, \tag{A.76}$$

where y is the rating given to an aspect of a music excerpt, $x_1, x_2, \ldots, x_p$ are variables for the excerpt under consideration, and $\alpha, \beta_1, \beta_2, \ldots, \beta_p$ are regression coefficients. Suppose that listeners are presented with already-discovered repeated patterns from one or more pieces of music, and asked to rate each pattern's musical importance on a scale from 1 (not at all important) to 10 (highly important). The rating given by a listener, which is more generally known as the *response*, is represented by y in (A.76). The variable x_1 could represent cardinality—the number of notes contained in one occurrence of a pattern. The next variable x_2 could represent the number of occurrences of a pattern in a particular excerpt, etc. Linear means linear in the coefficients, so linear models are a very broad family of functions. For instance, both of

$$\begin{aligned} \text{rating} &= \alpha + \beta_1 \text{cardinality} \cdot \text{occurrences}, & \text{(A.77)} \\ \text{rating} &= \alpha + \beta_1 \text{cardinality} + \beta_2 \text{occurrences}^2 & \text{(A.78)} \end{aligned}$$

are linear models.

In (A.76), the rating y is known, as are the values of the variables x_1, $x_2, \ldots$, x_p, so the aim is to estimate the coefficients $\theta = (\alpha, \beta_1, \beta_2, \ldots, \beta_p)$. This is done by considering a *linear regression model*

$$Y_i = \underbrace{\alpha + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p}}_{(\dagger)} + \varepsilon_i, \qquad i = 1, 2, \ldots, n. \tag{A.79}$$

Capital letters for the n ratings $Y_1, Y_2, \ldots, Y_n$ indicate that these are random variables. On the right hand side there is an expression (†) similar to that in (A.76). The notation has been altered so that $x_{i,j}$ is the value of the jth variable taken by the ith observation, where $i = 1, 2, \ldots, n$, and $j = 1, 2, \ldots, p$. These are often referred to as the *explanatory variables* or *predictors*, as they 'explain' the response (ratings). The terms $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are non-observable, assumed to be independent and normally distributed random variables (cf. Def. A.45) with zero mean and constant variance. Sometimes they are referred to as *departures*, as their inclusion in (A.79) adjusts for (†) 'departing' (being different) from Y_i. More commonly though, they are called *residual errors*. The coefficients $\theta = (\alpha, \beta_1, \beta_2, \ldots, \beta_p)$ are estimated so as to minimise the sum of squares of the departures, $\sum_{i=1}^n \varepsilon_i^2$.

Continuing with the example of rating already-discovered patterns, suppose that a listener rates the musical importance of five patterns as 9, 2, 8, 4, and 1, and that these patterns have respective cardinalities 15, 3, 4, 7, 3, and respective occurrences 3, 2, 5, 3, 2. This information can be expressed as

$$9 = \alpha + 15\beta_1 + 3\beta_2 + \varepsilon_1, \tag{A.80}$$

$$2 = \alpha + 3\beta_1 + 2\beta_2 + \varepsilon_2, \tag{A.81}$$

$$8 = \alpha + 4\beta_1 + 5\beta_2 + \varepsilon_3 \tag{A.82}$$

$$4 = \alpha + 7\beta_1 + 3\beta_2 + \varepsilon_4 \tag{A.83}$$

$$1 = \alpha + 3\beta_1 + 2\beta_2 + \varepsilon_5. \tag{A.84}$$

The ratings $y_1, y_2, \ldots, y_5$ are the observed values of the responses Y_1, $Y_2, \ldots,$ Y_5. The above simultaneous equations can be expressed more concisely as

$$\mathbf{y} = \mathbf{X}\theta + \varepsilon, \tag{A.85}$$

where

$$\mathbf{y} = \begin{pmatrix} 9 \\ 2 \\ 8 \\ 4 \\ 1 \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 15 & 3 \\ 1 & 3 & 2 \\ 1 & 4 & 5 \\ 1 & 7 & 3 \\ 1 & 3 & 2 \end{pmatrix}, \quad \theta = \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{pmatrix}. \tag{A.86}$$

The matrix $\mathbf{X}$ is often referred to as the *design matrix.* As mentioned above, the coefficients θ are estimated so as to minimise the sum of squares of the departures, $\sum_{i=1}^{n} \varepsilon_i^2$. The estimated coefficients are denoted with 'hats', $\widehat{\theta} = (\widehat{\alpha}, \widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_p)$. For this example (and more generally) it can be shown (but will not be shown here) that

$$\widehat{\theta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{A.87}$$

minimises $\sum_{i=1}^{n} \varepsilon_i^2$. Matrix transpose, multiplication, and inverses are defined in Def. A.3, and will be required to understand the following regression calculations:

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 15 & 3 & 4 & 7 & 3 \\ 3 & 2 & 5 & 3 & 2 \end{pmatrix} \begin{pmatrix} 1 & 15 & 3 \\ 1 & 3 & 2 \\ 1 & 4 & 5 \\ 1 & 7 & 3 \\ 1 & 3 & 2 \end{pmatrix} = \begin{pmatrix} 5 & 32 & 15 \\ 32 & 308 & 98 \\ 15 & 98 & 51 \end{pmatrix}, \tag{A.88}$$

$$(\mathbf{X}^T\mathbf{X})^{-1} \approx \begin{pmatrix} 1.984 & -0.053 & -0.482 \\ -0.053 & 0.010 & 0.003 \\ -0.482 & -0.003 & 0.168 \end{pmatrix}, \tag{A.89}$$

$$\mathbf{X}^T\mathbf{y} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 15 & 3 & 4 & 7 & 3 \\ 3 & 2 & 5 & 3 & 2 \end{pmatrix} \begin{pmatrix} 9 \\ 2 \\ 8 \\ 4 \\ 1 \end{pmatrix} = \begin{pmatrix} 24 \\ 204 \\ 85 \end{pmatrix}, \tag{A.90}$$

$$\begin{aligned} \widehat{\theta} &= \begin{pmatrix} \widehat{\alpha} \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ &\approx \begin{pmatrix} 1.984 & -0.053 & -0.482 \\ -0.053 & 0.010 & 0.003 \\ -0.482 & -0.003 & 0.168 \end{pmatrix} \begin{pmatrix} 24 \\ 204 \\ 85 \end{pmatrix} \approx \begin{pmatrix} -4.126 \\ 0.449 \\ 2.017 \end{pmatrix}. \end{aligned} \tag{A.91}$$

Therefore, in this example, the empirically derived formula for rating pattern importance is

$$\text{rating} = -4.126 + 0.449 \cdot \text{cardinality} + 2.017 \cdot \text{occurrences}. \tag{A.92}$$

The model on which this formula is based is flawed: too few data ($n = 5$); only two explanatory variables considered (cardinality and occurrences); no assumptions about $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_5$ were checked. It gives a flavour, however, for fitting a statistical model and deriving a formula empirically.

Example A.48. Analysis of variance (ANOVA). ANOVA is a special case of multiple linear regression (see above), where the predictors are binary variables that represent different *blocks* and/or *treatments* of observations. For instance, suppose that twelve people participate in a study on aural music skills, which is intended to investigate the efficacy of a new aural skills training method. Using a pre-training aural test, the participants are divided (by the median test score) into two blocks of six skilled and six unskilled participants. Within each block, the first two randomly selected participants undertake no aural skills training, the next two receive training according to an existing method called the Kodály Method (Choksy, 1974), and the last two participants receive training according to a new method called *Augment*. The different types of training are referred to as treatments. The performance of participants is assessed by a post-training aural test, and the response variable we are considering for each participant is labelled *improvement*: post-training test score minus pre-training test score. The design matrix for the regression will be

$$\begin{array}{c} \\ 1\\2\\3\\4\\5\\6\\7\\8\\9\\10\\11\\12 \end{array}\begin{array}{c} \begin{array}{cccc} \text{baseline} & \text{skilled} & \text{kodály} & \text{augment} \end{array} \\ \left(\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \end{array}\right) \end{array} = \mathbf{X}. \tag{A.93}$$

Regression calculations analogous to (A.87)-(A.91) can be performed to derive a formula for the improvement between post- and pre-training aural test scores, based on whether a participant was skilled or unskilled, and whether they received no training, training in the Kodály Method, or training in the Augment Method. The formula will be

$$\text{improvement} = \alpha + \beta_1 \cdot \text{skilled} + \beta_2 \cdot \text{kodály} + \beta_3 \cdot \text{augment}, \tag{A.94}$$

and so-called *contrasts*, such as $\beta_j - \beta_i$, will tell us about the effectiveness of one block or treatment over another. For example, if $\beta_3 - \beta_2$ is significantly greater than zero, this constitutes evidence that the new Augment Method is more effective for aural skills training than the Kodály Method. As with the previous example, this model is flawed due to too few data for the experimental design, and assumptions about $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{12}$ must be checked after the regression.

Multiple linear regression and ANOVA are treated in much more detail elsewhere (Lunn, 2007b; Daly et al., 1995; Davison, 2003).

References

Lois Choksy. *The Kodály method: comprehensive music education from infant to adult.* Prentice-Hall, Englewood Cliffs, NJ, 1974.

Tom Collins. *Improved methods for pattern discovery in music, with applications in automated stylistic composition.* PhD thesis, Faculty of Mathematics, Computing and Technology, The Open University, 2011.

Tom Collins, Jeremy Thurlow, Robin Laney, Alistair Willis, and Paul H. Garthwaite. A comparative evaluation of algorithms for discovering translational patterns in baroque keyboard works. In J. Stephen Downie and Remco Veltkamp, editors, *Proceedings of the International Society for Music Information Retrieval Conference*, pages 3–8, Utrecht, The Netherlands, 2010.

Tom Collins, Andreas Arzt, Sebastian Flossmann, and Gerhard Widmer. SIARCT-CFP: improving precision and the discovery of inexact musical patterns in point-set representations. In Alceu S. Britto Jr, Fabien Gouyon, and Simon Dixon, editors, *Proceedings of the International Society for Music Information Retrieval Conference*, pages 549–554, Curitiba, Brazil, 2013.

Fergus Daly, David J. Hand, M. Chris Jones, A. Daniel Lunn, and Kevin J. McConway. *Elements of statistics.* Addison-Wesley, Wokingham, UK, 1995.

Anthony C. Davison. *Statistical models.* Cambridge University Press, Cambridge, UK, 2003.

Daniel Lunn. Part A: Statistics lecture notes, 2007a. Retrieved 15 July, 2010 from http://www.stats.ox.ac.uk/∼dlunn/A5_06/A5_2006.htm.

Daniel Lunn. Part B: Applied statistics lecture notes, 2007b. Retrieved 15 July, 2010 from http://www.stats.ox.ac.uk/~dlunn/BS1_05 /BS1_mt05.htm.

Sheldon Ross. *A first course in probability.* Pearson Education, Inc, Upper Saddle River, NJ, 7th edition, 2006.

Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.