MATH1710 Probability and Statistics I

Matthew Aldridge

University of Leeds, 2021–22

Contents

Schedule **7**

About MATH1710 **9**

Organisation of MATH1710 9

Content of MATH1710 14

About these notes 15

Part I: EDA **19**

1 Exploratory data analysis **19**

1.1 What is EDA? 19

1.2 What is R? 20

1.3 Summary statistics and boxplots 21

1.4 Binned data and histograms 25

1.5 Multiple variables and scatterplots 28

Summary 30

Problem Sheet 1 **31**

A: Short questions 31

B: Long questions 32

C: Assessed questions 33

Solutions to short questions 34

Part II: Probability **37**

2 Probability spaces **37**

2.1 What is probability? 37

2.2 Sample spaces and events 38
2.3 Basic set theory 40
2.4 Probability axioms 44
2.5 Properties of probability 46
2.6 Addition rules for unions 47
Summary 49

3 Classical probability **51**
3.1 Probability with equally likely outcomes 51
3.2 Multiplication principle 52
3.3 Sampling with and without replacement 53
3.4 Ordering 54
3.5 Sampling without replacement in any order 56
3.6 Birthday problem 58
Summary 59

Problem Sheet 2 **61**
A: Short questions 61
B: Long questions 62
C: Assessed questions 63
Solutions to short questions 63

4 Independence and conditional probability **65**
4.1 Independent events 65
4.2 Conditional probability 66
4.3 Chain rule 68
4.4 Law of total probability 69
4.5 Bayes' theorem 70
4.6 Diagnostic testing 72
Summary 73

5 Discrete random variables **75**
5.1 What is a random variable? 75
5.2 Probability mass functions 76
5.3 Expectation 79

5.4 Functions of random variables 80
5.5 Variance 82
Summary 84

Problem Sheet 3 **85**
A: Short questions 85
B: Long questions 86
C: Assessed questions 87
Solutions to short questions 88

6 Discrete distributions **89**
6.1 Binomial distribution 89
6.2 Geometric distribution 91
6.3 Poisson distribution 93
6.4 Poisson approximation to the binomial 95
6.5 Distributions as models for data 98
Summary 99

7 Multiple random variables **101**
7.1 Joint distributions 101
7.2 Independence of random variables 103
7.3 Conditional distributions 104
7.4 Expectation sums and products 105
7.5 Covariance 107
7.6 Law of large numbers 109

Problem Sheet 4 **113**
A: Short questions 113
B: Long questions 114
C: Assessed questions 115
Solutions to short questions 116

Other stuff **119**

R Worksheets **119**

R worksheets 119

About R and RStudio 119

How to access R and RStudio 120

Installing R and RStudio 120

Troubleshooting drop-in sessions 121

Solutions **123**

Problem Sheet 1 123

Problem Sheet 2 129

Problem Sheet 3 135

Schedule

Week 7 (8–12 November):

- **Section 7:** Multiple random variables
- **Problem Sheet 4:** deadline for assessed questions: Monday 22 November
- **R Worksheet 7:** Discrete distributions deadline for assessed questions: Monday 15 November

Week 6 (1–5 November):

- **Section 6:** Discrete distributions
- **Problem Sheet 4:** deadline for assessed questions: Monday 22 November
- **R Worksheet 6:** R Markdown (optional)

Week 5 (25–29 October):

- **Section 5:** Discrete random variables
- **Problem Sheet 3:** deadline for assessed questions: Monday 8 November
- **R Worksheet 5:** Plots II: Making plots better deadline for assessed questions: Monday 1 November

Week 4 (18–22 October):

- **Section 4:** Independence and conditional probability
- **Problem Sheet 3:** deadline for assessed questions: Monday 8 November
- **R Worksheet 4:** Plots I: making plots

Week 3 (11–15 October):

- **Section 3:** Classical probability
- **Problem Sheet 2:** deadline for assessed questions: Monday 25 October
- **R Worksheet 3:** Data in R deadline for assessed questions: Monday 18 October

Week 2 (4–8 October):

- **Section 2:** Probability spaces
- **Problem Sheet 2:** deadline for assessed questions: Monday 25 October
- **R Worksheet 2:** Vectors

Week 1 (27 September – 1 October):

- **Section 1:** Exploratory data analysis
- **Problem Sheet 1:** deadline for assessed questions: Monday 11 October
- **R Worksheet 1:** R basics

About MATH1710

Organisation of MATH1710

This module is **MATH1710 Probability and Statistics I**. A few students will be taking this module as half of **MATH2700 Probability and Statistics for Scientists**.

This module lasts for 11 weeks from 27 September to 10 December 2021. The exam will take place between 10 and 21 January 2022.

The core teaching team are:

- Dr Matthew Aldridge (you can call me "Matt" or "Dr Aldridge"): I am the module leader, the main lecturer, and the main author of these notes.

The shared email address for the core teaching team is math1710@leeds.ac.uk; please use this address, rather than emailing our personal addresses; this will ensure your email is seen as soon as possible.

Notes and videos

The main way you will learn new material for this module is by reading these notes and by watching the accompanying pre-recorded videos. There will be one section of notes each week, for a total of 11 sections, with the final section being a summary and revision.

Reading mathematics is a slow process. Each section should take one and a half to two hours to work through; we recommend you split this into two or more sessions. If you find yourself regularly getting through sections in much less than that amount of time, you're probably not reading carefully enough through each sentence of explanation and each line of mathematics, including understanding the motivation, checking the accuracy, and making your own notes.

You are probably reading the web version of the notes. If you want a PDF or ebook copy (to read offline or to print out), they can be downloaded via the top ribbon of the page. (Warning: I have not made as much effort to make the PDF and ebook as neat and tidy as I have the web version, and there may be formatting errors.)

We are very keen to hear about errors in the notes mathematical, typographical or otherwise. Please, please email us if think you may have found any.

Problem sheets

There will be 5 problem sheets. Each problem sheet has a number of short and long questions for you to cover in your own time to help you learn the material, and two assessed questions, which you should submit for marking. The assessed questions on each problem sheet make up 3% of your mark on this module, for a total of 15%. Although the deadlines are on Mondays, you are advised to complete and submit the work in the previous week.

Problem Sheet	Sections covered	Assessed work due
1	1	Monday 11 October (Week 3)
2	2 and 3	Monday 25 October (Week 5)
3	4 and 5	Monday 8 November (Week 7)
4	6 and 7	Monday 22 November (Week 9)
5	8, 9 and 10	Monday 6 December (Week 11)

Assessed questions should be submitted in PDF format through Gradescope. (Further Gradescope details will follow.) Most students choose to hand-write their solutions and then scan them to PDF using their phone; you should use a proper scanning app – we recommend Microsoft Office Lens or Adobe Scan – and not just submit photographs.

Lectures

You will have one online synchronous (that is, live, not recorded) "lecture" session each week, with me, run through Zoom. Because this is a large cohort, we will split into two groups:

- Group 1: Mondays at 1200
- Group 2: Mondays at 1500

You should check your timetable to see which lecture group you are in.

This will not be a "lecture" in the traditional sense of the term, but will be an opportunity to re-emphasise material you have already learned from notes and videos, to give extra examples, and to answer common student questions, with some degree of interactivity via quizzes, polls, and the chat box.

We will assume you have completed all the work for the previous week by the time of the lecture.

We are very keen to hear about things you'd like to go through in the lectures; please email us with your suggestions.

Tutorials

Tutorials are small groups of about a dozen students. You have been assigned to one of 38 tutorial groups, each with a member of staff as the tutor. Your tutorial group will meet five times, in Weeks 2, 4, 6, 8, and 10. Tutorial groups will meet in person on campus; you should check your timetable to see when and where your tutorial group meets. (For those not yet on campus, due to travel restrictions or health conditions, there will be an extra online tutorial group for the first few tutorials.)

The main goal of the tutorials will be to go over your answers to the non-assessed questions on the problems sheets in an interactive session. In this smaller group, you will be able to ask detailed questions of your tutor, and have the chance to discuss your answers to the problem sheet. Your tutor may ask you to present some of your work to your fellow students, or may give you the opportunity to work together with others during the tutorial. Your tutor may be willing to give you a hint on the assessed questions if you've made a first attempt but have got stuck.

My recommended approach to problem sheets and tutorials is the following:

- Work through the problem sheet before the tutorial, spending plenty of time on it, and making multiple efforts at questions you get stuck on. I recommend spending *at least 3 hours per week* on the problem sheets, which will usually mean a total of *at least 6 hours per problem sheet* (as most problem sheets cover two weeks). Collaboration is encouraged when working through the non-assessed problems, but I recommend writing up your work on your own; answers to assessed questions must be solely your own work.
- Take advantage of the small group setting of the tutorial to ask for help or clarification on questions you weren't able to complete.
- After the tutorial, attempt again the questions you were previously stuck on.
- If you're still unable to complete a question after this second round of attempts, *then* consult the solutions.

Your tutor will also be the marker of your answers to the assessed questions on the problem sheets.

R worksheets

R is a programming language that is particularly good at working with probability and statistics. Learning to use R is an important part of this module, and is used in many other modules in the University, particularly in MATH1712 Probability and Statistics II. R is used by statisticians throughout academic and increasingly in industry too. Learning to program is a valuable skill for all students, and learning to use R is particularly valuable for students interested in statistics and related topics like actuarial science.

You will learn R by working through one R worksheet each week in your own time. Worksheets 3, 5, 7, 9 and 11 will also contain a couple of questions for assessment. Each of these is worth 3% of your mark for a total of 15%. I recommend spending one hour per week on the week's R worksheet, plus one extra hour if there are assessed questions that week.

You can read more about the language R, and about the program RStudio that we recommend you use to interact with R, in the R section of these notes.

To help you if you have problems with R, we have organised optional **R troubleshooting drop-in sessions**, where you can discuss any problems you have with an R expert, in Weeks 2 and 3. Check your timetable for details – these will be listed on your timetable as "practicals".

Optional "office hours" drop-in sessions

If you there is something in the module you wish to discuss privately one-on-one with the module core teaching team, the place for the is the optional weekly "office hours", which will operate as drop-in sessions. These sessions are an optional opportunity for you to ask questions you have to a member of staff; these are particularly useful if there's something on the module that you are stuck on or confused about, but we're happy to discuss any statistics-related issues or questions you have.

There will be two "office hours" drop-in sessions per week:

- Wednesday at 1000 in PRD 9.320 (Physics Research Deck)
- Wednesday at 1200 in Emmanuel Centre SR 02

(For timetabling reasons, the 1000 sessions appear on the timetable for MATH2700 students and the 1200 sessions appear on the timetable for MATH1710 students, but I'm happy for anyone to attend either hour.)

Time management

It is, of course, up to you how you choose to spend your time on this module. But my recommendations for your weekly work would be something like this:

- **Notes and videos:** 2 hours per week/section
- **Problem sheet:** 3 hours per week (so 6 hours for most problem sheets) plus 1 extra hour for writing up and submitting answers to assessed questions
- **R worksheet:** 1 hour per week/worksheet, plus 1 extra hour if there are assessed questions
- **Lecture:** 1 hour per week
- **Tutorial:** 1 hour every other week
- **Revision:** 13 hours total at the end of the module

That's roughly 8 hours a week, and makes 100 hours in total. (MATH1710 is a 10 credit module, so is supposed to represent 100 hours work. MATH2700 students are expected to be able to use their greater experience to get through the material in just 75 hours, so should scale these recommendations accordingly.)

Exam

There will be an exam in January, which makes up the remaining 70% of your mark. The exam will consist of 20 short and 2 long questions, and will be time-limited to 2 hours. We'll talk more about the exam format near the end of the module.

Who should I ask about...?

Remember that the email address for the core module teaching team is math1710@leeds.ac.uk. Please don't email our personal addresses; it will take longer for us to reply, and we may miss your email all together.

- *I don't understand something in the notes or on a problem sheet*: Come to office hours, or (if the timing works) ask your tutor in your next tutorial.
- *I'm having difficulties with R:* In Weeks 2 or 3, you should attend the R trouble-shooting drop-in session; at other times, come to office hours.
- *I have an admin question about arrangements for the module:* Come to office hours or email the core module teaching team.
- *I have an admin question about arrangements for my tutorial:* Contact your tutor.
- *I have an admin question about general arrangements for my course as a whole:* Email the Maths Taught Students Office (Maths.Taught.Students@leeds.ac.uk) or speak to your personal academic tutor.
- *I have a question about the marking of my assessed work on the problem sheets:* First, check your feedback on Gradescope; if you still have questions, contact your tutor.
- *I have a question about the marking of my assessed work on the R worksheets:* Come to office hours or email the core module teaching team.
- *I have suggestion for something to cover in the lectures:* Email the core module teaching team.
- *Due to exceptional personal circumstances I require an extension on or exemption from assessed work:* Email the Maths Taught Students Office; neither the core module teaching team nor your tutor are able to offer extensions or exemptions. (Only exemptions, not extensions, are available for R worksheets.)

Content of MATH1710

Prerequisites

The formal prerequisite for MATH1710 is "Grade B in A-level Mathematics or equivalent". We'll assume you have some basic school-level maths knowledge, but we don't assume you've studied probability or statistics in detail before (although we recognise that many of you will have). If you have studied probability and/or statistics at A-level (or post-16 equivalent) level, you'll recognise some of the material in this module; however you should find that we go deeper in some areas, and that we treat the material through with a greater deal of mathematical formality and rigour. "Rigour" here means precisely stating our assumptions, and carefully *proving* how other statements follow from those assumptions.

Syllabus

The module has three parts: a short first part on "exploratory data analysis", a long middle part on probability theory, and a short final part on a statistical framework called "Bayesian statistics". There's also the weekly R worksheets, which you could count as a fourth part running in parallel, but which will connect with the other parts too.

An outline plan of the topics covered is the following. (Remember that one section is one week's work.)

- **Exploratory data analysis** [1 section] Summary statistics, data visualisation
- **Probability** [8 sections]
  - Probability with events: Probability spaces, probability axioms, examples and properties of probability, "classical probability" of equally likely events, independence, conditional probability, Bayes' theorem [3 sections]
  - Probability with random variables: Discrete random variables, expectation and variance, binomial distribution, geometric distribution, Poisson distribution, multiple random variables, law of large numbers, continuous random variables, exponential distribution, normal distribution, central limit theorem [5 sections]
- **Bayesian statistics** [1 section]: Bayesian framework, Beta prior, normal–normal model
- Summary and revision [1 section]

Books

You can do well on this module by reading the notes and watching the videos, attending the lectures and tutorials, and working on the problem sheets and R

worksheets, without needing to do any further reading beyond this. However, students can benefit from optional extra background reading or an alternative view on the material, especially in the parts of the module on probability.

For exploratory data analysis, you can stick to Wikipedia, but if you really want a book, I'd recommend:

- GM Clarke and D Cooke, *A Basic Course in Statistics*, 5th edition, Edward Arnold, 2004.

For the probability section, any book with a title like "Introduction to Probability" would do. Some of my favourites are:

- JK Blitzstein and J Hwang, *Introduction to Probability*, 2nd edition, CRC Press, 2019.
- G Grimmett and D Welsh, *Probability: An Introduction*, 2nd edition, Oxford University Press, 2014. (The library has online access.)
- SM Ross, *A First Course in Probability*, 10th edition, Pearson, 2020.
- RL Scheaffer and LJ Young, *Introduction to Probability and Its Applications*, 3rd edition, Cengage, 2010.
- D Stirzaker, *Elementary Probability*, 2nd edition, Cambridge University Press, 2003. (The library has online access.)

I also found lecture notes by Prof Oliver Johnson (University of Bristol) and Prof Richard Weber (University of Cambridge) to be useful.

On Bayesian statistics, I recommend:

- JV Stone, *Bayes' Rule: A Tutorial Introduction to Bayesian Analysis*, Sebtel Press, 2013.

For R, there are many excellent resources online, and Google is your friend for finding them.

(For all these books I've listed the newest editions, but older editions are usually fine too.)

About these notes

These notes were written by Matthew Aldridge in 2021. They are based in part on previous notes by Dr Robert G Aykroyd and Prof Wally Gilks. Dr Jason Anquandah and Dr Aykroyd advised on the R worksheets. Dr Aykroyd's help and advice on many aspects of the module was particularly valuable.

These notes (in the web format) should be accessible by screenreaders. The videos have (highly imperfect) automated subtitles. If you have accessibility difficulties with these notes, contact maths1710@leeds.ac.uk.

Part I: EDA

Chapter 1

Exploratory data analysis

1.1 What is EDA?

Statistics is the study of data. **Exploratory data analysis** (or **EDA**, for short) is the part of statistics concerned with taking a "first look" at some data. Later, toward the end of this course, we will see more detailed and complex ways of building models for data, and in MATH1712 Probability and Statistics II (for those who take it) you will see many other statistical techniques – in particular, ways of testing formal hypotheses for data. But here we're just interested in first impressions and brief summaries.

In this section, we will concentrate on two aspects of EDA:

- **Summary statistics:** That is, calculating numbers that briefly summarise the data. A summary statistic might tell us what "central" or "typical" values of the data are, how spread out the data is, or about the relationship between two different variables.
- **Data visualisation:** Drawing a picture based on the data is an another way to show the shape (centrality and spread) of data, or the relationship between different variables.

Even before calculating summary statistics or drawing a plot, however, there are other questions it is important to ask about the data:

- *What is the data?* What variables have been measured? How were they measured? How many datapoints are there? What is the possible range of responses?
- *How was the data collected?* Was data collected on the whole population or just a smaller sample? (If a sample: How was that sample chosen? Is that sample representative of the population?) How were there variable measured?
- *Are there any outliers?* "Outliers" are datapoints that seem to be very different from the other datapoints – for example, are much larger or

much smaller than the others. Each outlier should be investigated to seek the reason for it. Perhaps it is a genuine-but-unusual datapoint (which is useful for understanding the extremes of the data), or perhaps there is an extraordinary explanation (a measurement or recording error, for example) meaning the data is not relevant. Once the reason for an outlier is understood, it then *might* be appropriate to exclude it from analysis (for example, the incorrectly recorded measurement). It's usually bad practice to exclude an outlier merely for being an outlier before understanding what caused it.

- *Ethical questions:* Was the data collected ethically and, where necessary, with the informed consent of the subjects? Has it been stored properly? Are their privacy issues with the collection and storage of the data? What ethical issues should be considered before publishing (or not publishing) results of the analysis? Should the data be kept confidential, or should it be openly shared with other researchers for the betterment of science?

1.2 What is R?

R is a programming language that is particularly good at working with probability and statistics. A convenient way to use the language R is through the program **RStudio**. An important part of this module is learning to use R, by completing weekly worksheets – you can read more in the R section of these notes.

R can easily and quickly perform all the calculations and draw all the plots in this section of notes on exploratory data analysis. In this text, we'll show the relevant R code. Code will appear like this:

```
data <- c(4, 7, 6, 7, 4, 5, 5)
mean(data)
```

```
## [1] 5.428571
```

Here, the code in the first shaded box is the R commands that are typed into RStudio, which you can type in next to the `>` arrow in the RStudio "console". The numerical answers that R returns are shown here in the second unshaded box next to a double hashsign `##`. The `[1]` can be ignored (this is just R's way of saying that this is the first part of the answer – but all the answers here only have one part anyway). Plots produced by R are displayed here as pictures.

Most importantly for now, *you are not expected to understand the R code in this section yet.* The code is included so that, in the future, as you work through the R worksheets week by week, you can look back at the code in the section, and it will start to make sense. By the time you have finished R Worksheet 5 in week 5, you should be able understand most of the R code in this section.

1.3 Summary statistics and boxplots

Suppose we have collected some data on a certain variable. We will assume here that we have n datapoints, each of which is a single real number. We can write this data as a vector

$$\mathbf{x} = (x_1, x_2, \dots, x_n).$$

A **statistic** is a calculation from the data $\mathbf{x}$, which is (usually) also a real number. In this section we will look at two types of "summary statistics", which are statistics that we feel will give us useful information about the data.

We'll look here at two types of summary statistic:

- **Measures of centrality**, which tell us where the "middle" of the data is.
- **Measures of spread**, which tell us how far the data typically spreads out from that middle.

Some measures of centrality are the following.

Definition 1.1. Consider some real-valued data $\mathbf{x} = (x_1, x_2, \dots, x_n)$.

- The **mode** is the most common value of x_i. (If there are multiple joint-most common values, they are all modes.)
- Suppose the data is ordered as $x_1 \leq x_2 \leq \cdots \leq x_n$. Then the **median** is the central value in the ordered list. If n is odd, this is $x_{(n+1)/2}$; if n is even, we normally take halfway between the two central points, $\frac{1}{2}(x_{n/2}+x_{n/2+1})$.
- The **mean** $\bar{x}$ is

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

(In that last expression, we've made use of Sigma notation to write down the sum.)

Example 1.1. Some packets of Skittles (a small fruit-flavoured sweet) were opened, and the number of Skittles in each packet counted. There were 13 packets, and the number of sweets (sorted from smallest to largest) were:

$$59,\ 59,\ 59,\ 59,\ 60,\ 60,\ 60,\ 61,\ 62,\ 62,\ 62,\ 63,\ 63.$$

The mode is 59, because there were 4 packets containing 59 sweets; more than any other number. Since there are $n = 13$ packets, the middle packet is number $i = 7$, so the median is $x_7 = 60$. The mean is

$$\bar{x} = \frac{1}{13}(59 + 59 + \cdots + 63) = \frac{789}{13} = 60.7.$$

The median is one example of a "quantile" of the data. Suppose our data is increasing order again. For $0 \leq \alpha \leq 1$, the **α-quantile** $q(\alpha)$ of the data is the datapoint α of the way along the list. So the median is the $\frac{1}{2}$-quantile $q(\frac{1}{2})$, the minimum is the 0-quantile $q(0)$, and the maximum is the 1-quantile $q(1)$. Generally, $q(\alpha)$ is equal to $x_{1+\alpha(n-1)}$ when $1+\alpha(n-1)$ is an integer. (If $1+\alpha(n-1)$ isn't an integer, there are various conventions of how to choose that we won't go into here. R has *seven* different settings for choosing quantiles! – we will always just use R's default choice.)

Two other common terms: $q(\frac{3}{4})$ is called the **upper quartile** and $q(\frac{1}{4})$ is called the **lower quartile** (note "quartile" – as in "quarter" – not "quantile", here). The upper and lower quartiles of the $n = 13$ Skittles packets are the $q(\frac{1}{4}) = x_4 = 59$ and $x_{10} = 62$.

Some measures of spread are:

Definition 1.2. The **number of distinct observations** is precisely that: the number of different datapoints you have after removing any repeats.

The **interquartile range** is the difference between the upper and lower quartiles $\text{IQR} = q(\frac{3}{4}) - q(\frac{1}{4})$.

The **sample variance** is

$$s_x^2 = \frac{1}{n-1}\left((x_1-\bar{x})^2 + \cdots + (x_n-\bar{x})^2\right) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2,$$

where $\bar{x}$ is the sample mean from before. The **standard deviation** $s_x = \sqrt{s_x^2}$ is the square-root of the sample variance.

The formula we've given for sample variance is sometimes called the "definitional formula", as it's the formula used to *define* the sample variance. We can rearrange that formula as follows:

$$\begin{aligned}
s_x^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2x_i\bar{x} + \sum_{i=1}^{n}\bar{x}^2\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \bar{x}^2\sum_{i=1}^{n} 1\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right).
\end{aligned}$$

Here, the first line is the definitional formula; the second line is from expanding out the bracket; the third line is taking the sum term-by-term; the fourth line

takes any constants (things not involving i) outside the sums; the fifth line uses $\sum_{i=1}^{n} x_i = n\bar{x}$, from the definition of the mean, and $\sum_{i=1}^{n} 1 = 1 + 1 + \cdots 1 = n$; and the sixth line simplifies the final two terms.

This has left us with

$$s_x^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right).$$

This is sometimes called the "computational formula"; this is because it's usually more convenient to calculate the sample variance using this formula rather than the definitional formula.

The following R code reads in some data which has the daily average temperature in Leeds in 2020, divided into months. We can find, for example, the mean October temperature or the sample variance of the July temperature.

```
temperature <- read.csv("https://mpaldridge.github.io/math1710/data/temperature.csv")
jul <- temperature[temperature$month == "jul", ]
oct <- temperature[temperature$month == "oct", ]

mean(oct$temp)
```

```
## [1] 11.93548
```

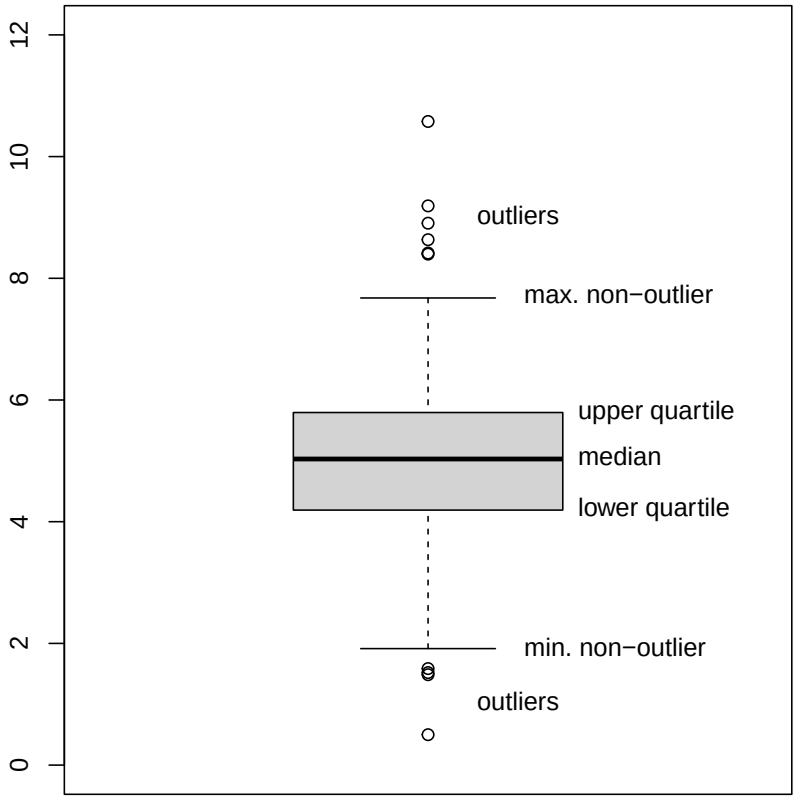
```
var(jul$temp)
```

```
## [1] 12.03226
```

A **boxplot** is a useful way to illustrate data. It can be easier to tell the difference between different data sets "by eye" when looking at a boxplot, rather than examining raw summary statistics.

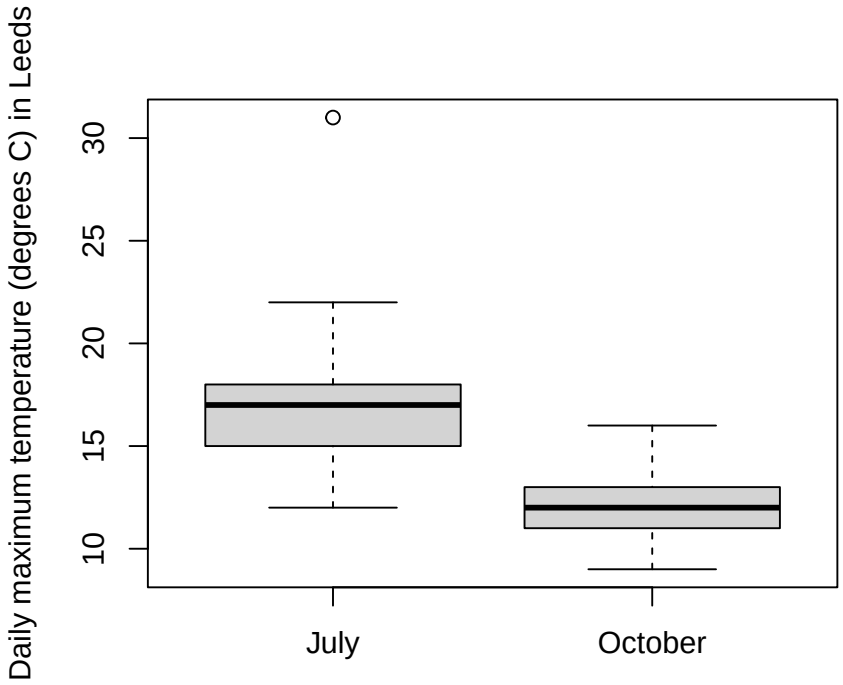
A boxplot is drawn as follows:

- The vertical axis represents the data values.
- Draw a box from the lower quartile $q(\frac{1}{4})$ to the median $q(\frac{1}{2})$.
- Draw another box on top of this from the median $q(\frac{1}{2})$ to the upper quartile $q(\frac{3}{4})$. Note that size of these two boxes put together is the interquartile range.
- Decide which datapoints are outliers, and plot these with circles. (The R default is that any data point less than $q(\frac{1}{4}) - 1.5 \times \text{IQR}$ or greater than $q(\frac{3}{4}) + 1.5 \times \text{IQR}$ is an outlier.)
- Out from the two previous boxes, draw "whiskers" to the smallest and largest non-outlier datapoints.



Here are two boxplots from the July and October temperature data. What do you conclude about the data from these boxplots?

```
boxplot(jul$temp, oct$temp,
        names = c("July", "October"),
        ylab = "Daily maximum temperature (degrees C) in Leeds")
```



(And yes, I did check the outlier to make sure it was a genuine datapoint.)

1.4 Binned data and histograms

Often when collecting data, we don't collect exact data, but rather collect data clumped into "bins". For example, suppose a student wished to use a questionnaire to collect data on how long it takes people to reach campus from home; they might not ask "Exactly how long does it take?", but rather give a choice of tick boxes: "0–5 minutes", "5–10 minutes", and so on.

Consider the following binned data, from $n = 100$ students:

Time	Frequency	Relative frequency
0–5 minutes	4	0.04
5–10 minutes	8	0.08
10–15 minutes	21	0.21
15–30 minutes	42	0.42
30–45 minutes	15	0.15
45–60 minutes	8	0.08
60–120 minutes	2	0.02

Time	Frequency	Relative frequency
Total	100	1

Here the **frequency** f_j of bin j is simply the number of observations in that bin; so, for example, 42 students had journey lengths of between 15 and 30 minutes. The **relative frequency** of bin j is f_j/n; that is, the proportion of the observations in that bin.

What is the median journey length? Well, we don't know exactly, but $0.04 + 0.08 + 0.21$ (the first three bins) is less than 0.5, while $0.04 + 0.08 + 0.21 + 0.42$ (including the fourth bin) is greater than 0.5. So we know that the median student is in the fourth bin, the "15–30 minute" bin, and we can say that the median journey length is between 15 and 30 minutes.

What about the mode? The bin with the most observations in it is the "15–30 minute" bin. But this bin covers 15 minutes, while some of the other bins only cover 5 minutes. It would be a fairer comparison to look at the **frequency density**: the relative frequency divided by the size of the bin.

Time	Frequency	Relative frequency	Frequency density
0–5 minutes	4	0.04	0.008
5–10 minutes	8	0.08	0.016
10–15 minutes	21	0.21	0.042
15–30 minutes	42	0.42	0.028
30–45 minutes	15	0.15	0.010
45–60 minutes	8	0.08	0.005
60–120 minutes	2	0.02	0.0003
Total	100	1	

In the first row, for example, the relative frequency is 0.04 and the size of the bin is 5 minutes, so the frequency density is $0.04/5 = 0.008$. So the modal bin – the bin with the highest frequency *density* – is in fact the "10–15 minutes" bin.

Since we don't have the exact data, it's not possible to exactly calculate the mean and variance. However, we can often get a good estimate by assuming that each observation was in fact right in the centre of its bin. So, for example, we can assume that all 4 observations in the "0–5 minutes" bin were journeys of exactly 2.5 minutes. Of course, this isn't true (or is highly unlikely to be true), but we can often get a good approximation this way.

For our journey-time data, our approximation of the mean would be

$$\bar{x} = \frac{1}{100}(4 \times 2.5 + 8 \times 7.5 + \cdots + 2 \times 90) = 24.4.$$

More generally, if m_j is the midpoint of bin j and f_j its frequency, then we can

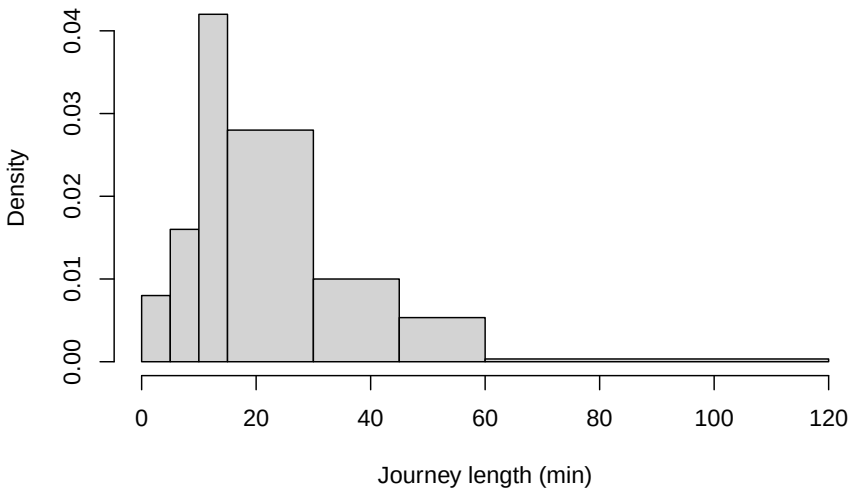
calculate the binned mean and binned variance by

$$\bar{x} = \frac{1}{n}\sum_j f_j m_j$$
$$s_x^2 = \frac{1}{n-1}\sum_j f_j (m_j - \bar{x})^2$$

Data in bins can be illustrated with a **histogram**. A histogram has the measurement on the x-axis, with one bar across the width of each bin, with bars drawn up to the height of the corresponding frequency density. Note that this means that the area of the bar is exactly the relative frequency of the corresponding bin. (If all the bins are the same width, frequency density is directly proportional to frequency and to relative frequency, so it can be clearer use one of those as the y-axis instead.)

Here is a histogram for our journey-time data:

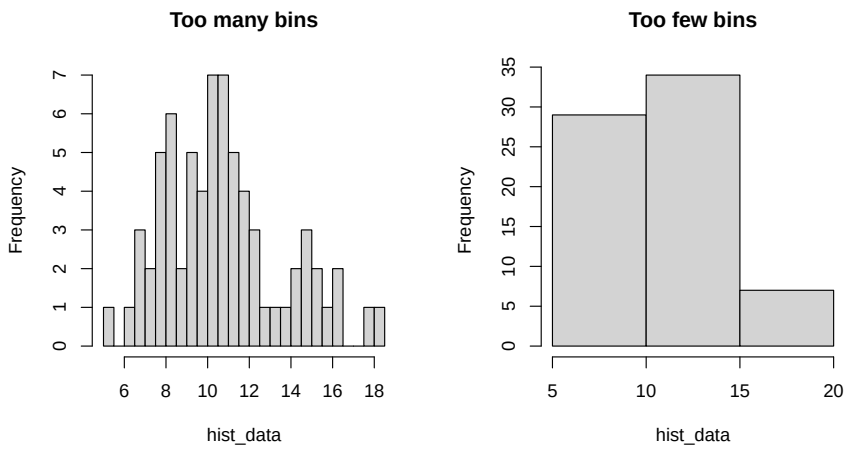
```
journeys <- read.csv("https://mpaldridge.github.io/math1710/data/journeys.csv")
bins <- c(0, 5, 10, 15, 30, 45, 60, 120)

hist(journeys$midpoint, breaks = bins,
     xlab = "Journey length (min)",
     main = "")
```



Often we draw histograms because the data was collected in bins. But even when we have exact data, we might choose to divide it into bins for the purposes of drawing a histogram. In this case we have to decide where to put the "breaks" between the bins. Too many breaks too close together, and the small number of

observations in each bin will give "noisy" results (see left); too few breaks too far apart, and the histogram will lose detail (see right).

```
hist_data <- c(rnorm(30, 8, 2), rnorm(40, 12, 3))  # Some fake data
hist(hist_data, breaks = 40, main = "Too many bins")
hist(hist_data, breaks = 3,  main = "Too few bins")
```



1.5 Multiple variables and scatterplots

Often, more than one piece of data is collected from each subject, and we wish to compare that data, to see if there is a relationship between the variables.

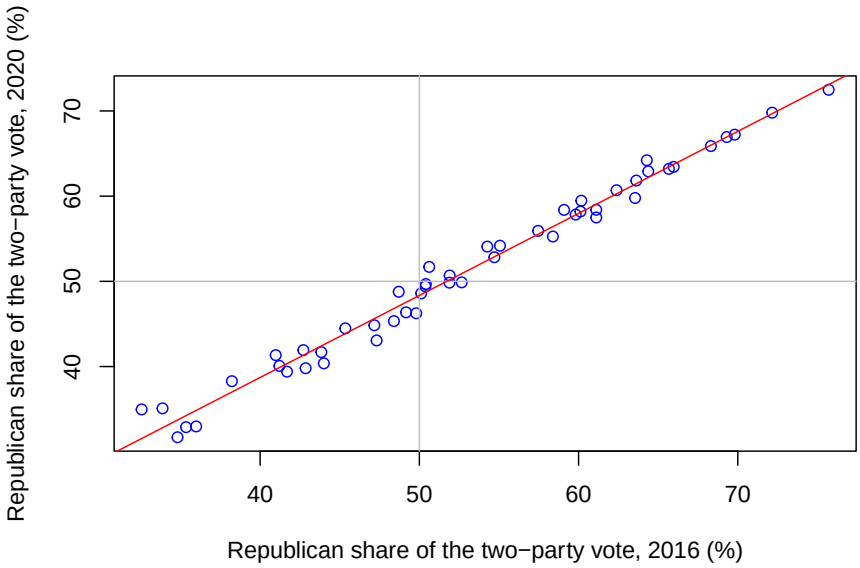
For example, we could take n second-year maths students, and for each student i, collect their mark x_i in MATH1710 and their mark y_i in MATH1712. This gives is two "paired" datasets, $\mathbf{x} = (x_1, x_2, \dots, x_n)$ and $\mathbf{y} = (y_1, y_2, \dots, y_n)$. We can calculate sample statistics of draw plots for $\mathbf{x}$ and for $\mathbf{y}$ individually. But we might also want to see if there is a relationship *between* $\mathbf{x}$ and $\mathbf{y}$: Do students with high marks in MATH1710 also get high marks in MATH1712?

A good way to visualise the relationship between two variables is to use a **scatterplot**. In a scatterplot, the ith data pair (x_i, y_i) is illustrated with a mark (such as a circle or cross) whose x-coordinate has the value x_i and whose y-coordinate has the value y_i.

In the following scatterplot, we have $n = 50$ datapoints for the 50 US states; for each state i, x_i is the Republican share of the vote in that state in the 2016 Trump–Clinton presidential election, and y_i is the Republican share of the vote in that state in the 2020 Trump–Biden election.

```
elections <- read.csv("https://mpaldridge.github.io/math1710/data/elections.csv")

plot(elections$X2016, elections$X2020,
```

```
     col = "blue",
     xlab = "Republican share of the two-party vote, 2016 (%)",
     ylab = "Republican share of the two-party vote, 2020 (%)")

abline(h = 50, col = "grey")
abline(v = 50, col = "grey")
abline(0.195, 0.963, col = "red")
```



We see that there is a strong relationship between **x** and **y**, with high values of x corresponding to high values of y and vice versa. Further, the points on the scatterplot lie very close to a straight line.

A useful summary statistic here is the **correlation**

$$r_{xy} = \frac{s_{xy}}{s_x s_y},$$

where s_{xy} is the **sample covariance**

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}),$$

and s_x and s_y are the standard deviations.

The correlation r_{xy} is always between -1 and $+1$. Values of r_{xy} near $+1$ indicate that the scatterpoints are close to a straight line with an upward slope (big x = big y); values of r_{xy} near -1 indicate that the scatterpoints are close to a straight line with a downward slope (big x = small y); and values of r_{xy} near 0 indicate that there is a weak linear relationship between x and y.

For the elections data, the correlation is

```
cor(elections$X2016, elections$X2020)
```

```
## [1] 0.9919659
```

which, as we expected, is extremely high.

Summary

- Exploratory data analysis is about taking a first look at data.
- Summary statistics are numbers calculated from data that give us useful information about the data.
- Summary statistics that measure the centre of the data include the mode, median, and mean.
- Summary statistics that measure the spread of the data include the number of distinct outcomes, the interquartile range, and the sample variance.
- A summary statistic that measures the linear relationship between two variables is the correlation.
- Boxplots, histograms, and scatterplots are useful ways of visualising data.

Problem Sheet 1

This is Problem Sheet 1, which covers material from Section 1 of the notes. You should work through all the questions on this problem sheet during Week 1, in preparation for your tutorial in Week 2. Questions C1 and C2 are assessed questions, and are due in by **2pm on Monday 11 October**. I recommend spending about 3 hours on this problem sheet in Week 1, plus 1 extra hour in Week 2 to neatly write up and submit your answers to the assessed questions.

A: Short questions

The first two questions are **short questions**, which are intended to be mostly not too difficult. Short questions usually follow directly from the material in the notes. Here, your should clearly state your final answer, and give enough working-out (or a short written explanation) for it to be clear how you reached that answer. You can check your answers with the solutions-without-working at the bottom of this sheet; solutions-with-working will be available later. If you get stuck on any of these questions, you might want to ask for guidance in your tutorial.

A1. Consider again the "number of Skittles in each packet" data from Example 1.1.

$$59,\ 59,\ 59,\ 59,\ 60,\ 60,\ 60,\ 61,\ 62,\ 62,\ 62,\ 63,\ 63.$$

(a) Calculate the mean number of Skittles in each packet.

(b) Calculate the sample variance using the computational formula.

(c) Calculate the sample variance using the definitional formula.

(d) Out of (b) and (c), which calculation did you find easier, and why?

A2. Consider the following data sets of the age of elected politicians on a local council. (The "18–30" consists of people older than and including 18, and younger than but *not* including 30.)

Age (years)	Frequency	Relative frequency	Frequency density
18–30	1		
30–40	3		
40–45	4		

Age (years)	Frequency	Relative frequency	Frequency density
45–50	5		
50–55	3		
55–60	1		
60–70	3		
Total	20	1	—

(a) Complete the table by filling in the relative frequency and frequency densities.

(b) What is the median age bin?

(c) Calculate (an approximation of) the mean age of the politicians.

B: Long questions

The next four questions are **long questions**, which are intended to be harder. Long questions often require you to think originally for yourself, not just directly follow procedures from the notes. Here, your answers should be written in complete sentences, and you should carefully explain in words each step of your working. Your answers to these questions – not only their mathematical content, but also how to clearly write good solutions – are likely to be the main topic for discussion in your tutorial.

B1. For each of the two datasets below, calculate the following summary statistics, or explain why it is not possible to do so: mode; median; mean; number of distinct outcomes; inter-quartile range; and sample variance.

(a) Six packets of Skittles are opened together, and the total number of sweets of each colour is:

Colour	Red	Orange	Yellow	Green	Purple
Number of Skittles	67	71	87	74	62

(b) Shirt sizes for a university football squad:

Colour	Xtra Small	Small	Medium	Large	Xtra Large
Number of shirts	0	1	6	4	5

*[**Note:** This has been corrected from an earlier version, where the 4 Large and 5 Xtra Large were the wrong way round.]*

B2. A summary statistic is informally said to be "robust" if it typically doesn't change much if a small number of outliers are introduced to a large dataset, or "sensitive" if it often changes a lot when a small number of outliers are introduced. Briefly discuss the robustness or sensitivity of the following summary

statistics: **(a)** mode; **(b)** median; **(c)** mean; **(d)** number of distinct outcomes; **(e)** inter-quartile range; and **(f)** sample variance.

B3. Let $\mathbf{a} = (a_1, a_2, \dots a_n)$ and $\mathbf{b} = (b_1, b_2, \dots, b_n)$ be two real-valued vectors of the same length. Then the *Cauchy–Schwarz inequality* says that

$$\left(\sum_{i=1}^{n} a_i b_i\right)^2 \leq \left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right).$$

Use the Cauchy–Schwarz inequality to show that the correlation r_{xy} satisfies $-1 \leq r_{xy} \leq 1$.

(*Hint:* Try to prove that $s_{xy}^2 \leq s_x^2 s_y^2$. How does this help?)

B4. A researcher wishes to study the effect of mental health on academic achievement. The researcher will collect data on the mental health of a cohort of students by asking them to fill in a questionnaire, and will measure academic achievement via the students' scores on their university exams. Discuss some of the ethical issues associated with the collection, storage, and analysis of this data, and with the publication of the results of the analysis. Are there ways to mitigate these issues?

(It's not necessary to write an essay for this question – a few short bulletpoints will suffice. There may be an opportunity to discuss these issues in more detail in your tutorial.)

C: Assessed questions

The last two questions are **assessed questions**. This means you will submit your answers, and your answers will be marked by your tutor. These two questions count for 3% of your final mark for this module. If you get stuck, your tutor may be willing to give you a hint in your tutorial.

The deadline for submitting your solutions is **2pm on Monday 11 October** at the beginning of Week 3, although I strongly recommend completing and submitting your work during Week 2. Submission will be via Gradescope; submission will open on Monday 4 October. You should submit your answers as a single PDF file. Most students choose to hand-write their work, then scan it to PDF using their phone; if you do this, you should use a proper scanning app (like Microsoft Lens or Adobe Scan) – please do not just submit photographs. We will discuss Gradescope submission further in the Week 2 lectures. Your work will be marked by your tutor and returned on Monday 18 September, when solutions will also be made available.

Question C1 is a "short question", where brief explanations or working are sufficient; Question C2 is a "long question", where the marks are not only for mathematical accuracy but also for the clarity and completeness of your explanations.

You should not collaborate with others on the assessed questions: your answers must represent solely your own work. The University's rules on academic integrity – and the related punishments for violating them – apply to your work on the assessed questions.

C1. The monthly average exchange rate for US dollars into British pounds over a 12-month period was:

$$1.306,\ 1.301,\ 1.290,\ 1.266,\ 1.290,\ 1.302,$$
$$1.317,\ 1.304,\ 1.284,\ 1.268,\ 1.247,\ 1.215.$$

(a) Calculate the median for this data.

(b) Calculate the mean for this data.

(c) Calculate the sample variance for this data.

(d) Is the mode an appropriate summary statistic for this data? Why/why not?

C2. **(a)** Prove the following computational formula for the sample covariance:

$$s_{xy} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right).$$

(b) Suppose that a dataset $\mathbf{x} = (x_1, x_2, \dots, x_n)$ (with $n \geq 2$) has sample variance $s_x^2 = 0$. Show that all the datapoints are in fact equal.

Solutions to short questions

A1. (a) 60.7 (b) & (c) 2.40 (d) — **A2.** (a) — (b) 45–50 (c) 47.3

Part II: Probability

Chapter 2

Probability spaces

2.1 What is probability?

We now begin the big central block of this module, on probability theory.

Probability theory is the study of randomness. Probability, as an area of mathematics, is a fascinating subject in its own right. However, probability is particularly important due to its usefulness in applications – especially in statistics (the study of data), in finance, and in actuarial science (the study of insurance).

Probability is well suited to modelling situations that involve randomness, uncertainty, or unpredictability. If we you want to predict the time of the next solar eclipse, a deterministic (that is, non-random) model based on physical laws will tell you when the sun, the moon, and the earth will be in the correct positions; but if you want to predict the weather tomorrow, or the price of a share of Apple stock next month, or the results of an election next year, you will need a probabilistic model that takes into account the uncertainty in the outcome. A probabilistic model could tell you the most likely outcome, or a range of the most probable outcomes.

So what do we mean when we talk about the "probability" of an event occurring? You might say that the probability of an event is a measure of "how likely" it is to occur, or what the "chance" of it occurring is.

More concretely, here are some interpretations of probability:

- **Subjective** (or **Bayesian**) **probability:** The probability of an event is the way someone expresses their degree of belief that the event will occur, based on their own judgement, and given the evidence they have seen. Their belief is measured on a scale from 0 to 1, from probabilities near 0 meaning they believe the event is very unlikely to occur to probabilities near 1 meaning they believe the event is very likely to occur.
  - This interpretation is philosophically sound, but a bit vague to be the basis for a mathematics module.

- **Classical** (or **enumerative**) **probability:** Suppose there are a finite number of equally likely outcomes. Then the probability of an event is the proportion of those outcomes that correspond to the event occurring. So when we say that a randomly dealt card has a probability $\frac{1}{13}$ of being an ace, this is because there are 52 cards of which 4 are aces, so the proportion of favourable outcomes is $\frac{4}{52} = \frac{1}{13}$.
  - This interpretation is good for simple procedures like flipping a fair coin, rolling a dice, or dealing cards, where the "finite number of equally likely outcomes" assumption holds. But we want to be able to study more complicated situations, where some outcomes are more likely than others, or where infinitely many different outcomes are possible.
- **Frequentist probability:** In a repeated experiment, the probability of an event is its long-run frequency. That is, if we repeat an experiment a very large number of times, the probability of the event is (approximately) the proportion of the experiments in which the event occurs. So when we say a biased coin has probability 0.9 of landing heads, we mean that were we toss it 1000 times, we would expect to see very close to $0.9 \times 1000 = 900$ heads.
  - There are two problems with this. First, this doesn't deal with events that can't be repeated over and over again (like "What's the probability that England win the 2022 World Cup?"). Second, to answer the question, "Yes, but *how* close to the probability should the proportion of occurrences be?", you end up having to answer, "Well, it depends on the probability," and you've got a circular definition.
- **Mathematical probability:** We have a function that assigns to each event a number between 0 and 1, called its probability, and that function has to obey certain mathematical rules, called "axioms".

It will not surprise you to learn that, in this mathematics course, we will take the "mathematical probability" approach. However, we will also learn useful things about the other approaches: we will see that classical probability is one special case of mathematical probability; we will see a result called the "law of large numbers" that says that the long-run frequency does indeed get closer and closer to the mathematical probability; and a result called "Bayes' theorem" will advise a subjectivist on how to update her subjective beliefs when she sees new evidence.

2.2 Sample spaces and events

Taking the "mathematical probability" approach, we will want to give a formal mathematical definition of the *probability* of an event. But even before that, we need to give a formal mathematical definition of an *event* itself. Our setup will be this:

- There is a set called the **sample space**, normally given the letter Ω (upper-case Omega), which is the set of all possible outcomes.

- An element of the sample space Ω is a **sample outcome**, sometimes given the letter ω (lower-case omega), represents one of the possible outcomes.
- An **event** is a set of sample outcomes; that is, a subset of the sample space Ω. Events are often given letters like A, B, C. We write $A \subset \Omega$ to mean that A is an event in (or, equivalently, is a subset of) the sample space Ω.

This will be easier to understand with some concrete examples. We write a set (such as a sample space or an event) by writing all the elements of that set inside curly brackets { }, separated by commas.

Example 2.1. Suppose we toss a (possibly biased) coin, and record whether it lands heads or tails. Then our sample space is $\Omega = \{\mathrm{H}, \mathrm{T}\}$, where the sample outcome H denotes heads and the sample outcome T denotes tails.

The event that the coin lands heads is $\{\mathrm{H}\}$.

Example 2.2. Suppose we roll a dice, and record the number rolled. Then our sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$, where the sample outcome 1 corresponds to rolling a one, and so on.

The event "we roll an even number" is $\{2, 4, 6\}$. The event "we roll at least a five" is $\{5, 6\}$.

Example 2.3. Suppose we wish to count how many claims are made to an insurance company in a year. We could model this by taking the sample space Ω to be $\mathbb{Z}_+ = \{0, 1, 2, \dots\}$, the set of all non-negative integers.

The event "the company receives less than 1000 claims" is $\{0, 1, 2, \dots, 998, 999\}$.

Example 2.4. Suppose we want a computer to pick a random number between 0 and 1. We could model this by taking the sample space Ω to be the interval $[0, 1]$ of all real numbers between 0 and 1.

The event "the number is bigger than $\frac{1}{2}$" is the sub-interval $(\frac{1}{2}, 1]$ of all real numbers greater than $\frac{1}{2}$ but no bigger than 1. The event "the first digit is a 7" is the sub-interval $[0.7, 0.8)$. The event "the random number is exactly $1/\sqrt{2}$" is $\{1/\sqrt{2}\}$.

In the first two examples, the sample space Ω was finite. In third example, the sample space was infinite but "countably infinite", in that it could be counted using the discrete values of the positive integers. Both of these were for *counting* discrete observations. In the fourth example, the sample space was infinite but "uncountably infinite", in that it had a sliding scale or "continuum" of gradually varying measurements. This was for *measuring* continuous observations. This distinction will be important later in the course.

For any sample space Ω, there are two special events that always exist. There's Ω itself, the event containing all of the sample outcomes, which represents "something happens". There's also the empty set $\emptyset$, which contains none of the sample outcomes, which represents "nothing happens". Common sense suggests that Ω should have probability 1, because *something* is bound to happen – this will later be one of our probability "axioms". Common sense also suggests that $\emptyset$

should have probability 0, because it can't be that *nothing* happens – this will not be one probability axioms, but we'll show that it follows logically from the axioms we do choose.

2.3 Basic set theory

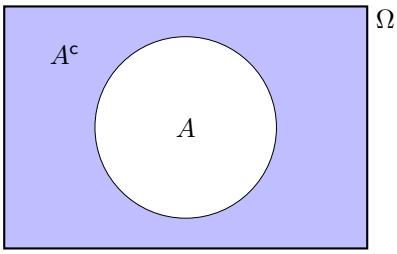
Since we've now defined events as being sets – specifically, subsets of the sample space Ω – it will be useful to mention a little set theory here.

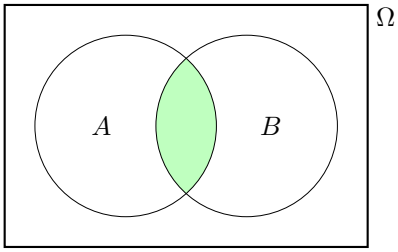
First, there are ways we can build new sets (or events) out of old. It's fine to just read the words and look at the pictures for these definitions, but those who want to read the equations too will need to know this:

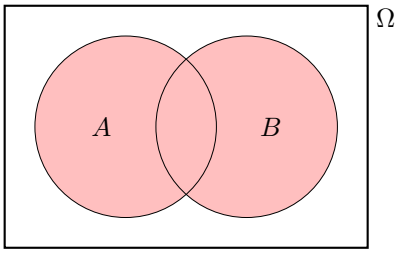
- $\omega \in A$ means "ω is in A" or "ω is an element of A", while $\omega \notin A$ means the opposite: ω is *not* in A;
- a colon : in the middle of set notation should be read as "such that";
- so $\{\omega \in \Omega : \text{fact about } \omega\}$ should be read as "the set of sample points ω in the sample space Ω such that the fact is true".

Definition 2.1. Consider a sample space Ω, and let A and B be events in that sample space.

- **NOT:** The **complement** of A, written A^{c} (and said "A complement" or "not A"), is the set of sample points not in A; that is
$$A^{\mathsf{c}} = \{\omega \in \Omega : \omega \notin A\}.$$
This represents the event that A does not occur.
- **AND:** The **intersection** of A and B, written $A \cap B$ (and said "A intersect B" or "A and B") is the set of sample points in both A and B; that is,
$$A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}.$$
This represents the event that both A and B occur.
- **OR:** The **union** of A and B, written $A \cup B$ (and said "A union B" or "A or B") is the set of sample points in A or in B; that is,
$$A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\}.$$
This represents the event that A occurs or B occurs. (In mathematics, "or" includes "both", so a sample outcome in both A and B is in $A \cup B$ too.)







Example 2.5. Suppose we are rolling a dice, so our sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$. Let $A = \{2, 4, 6\}$ be the event that we roll and even number, and let $B = \{5, 6\}$ be the event that we roll at least a 5. Then

$$\begin{aligned} A^{\mathsf{c}} &= \{1, 3, 5\} = \{\text{roll an odd number}\}, \\ A \cap B &= \{6\} = \{\text{roll a 6}\}, \\ A \cup B &= \{2, 4, 5, 6\}. \end{aligned}$$

An important case is when two events A, B cannot happen at the same time; that is, $A \cap B = \emptyset$ ("A intersect B is the empty set"). In this case, we say that A and B are **disjoint** or **mutually exclusive**. For example, when Ω is a deck of cards, then $A = \{\text{the card is a spade}\}$ and $B = \{\text{the card is red}\}$ are disjoint, because a card cannot be both a spade (a black suit) and red.

There are a few rules about combining the complement, intersection and union operations.

- The **double complement law** tells us that not-not-A is the same as A:

  $$(A^{\mathsf{c}})^{\mathsf{c}} = A.$$

  This says that if it's not "not-raining", then it's raining!
- The **distributive laws** tells us we can "mutiply out of the brackets brackets" with sets:

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

- **De Morgan's laws** tell us how complements interact with intersection/unions:

  $$\begin{aligned} (A \cap B)^{\mathsf{c}} &= A^{\mathsf{c}} \cup B^{\mathsf{c}} \\ (A \cup B)^{\mathsf{c}} &= A^{\mathsf{c}} \cap B^{\mathsf{c}} \end{aligned}$$

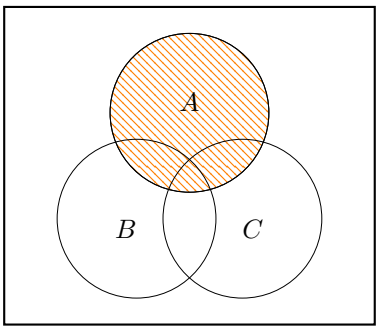
The first of these says that if it's not a Monday in October, then either it's not Monday or it's not October (or both). The second says that if a maths lecture is not "useful or fun", then it's not useful and it's not fun.

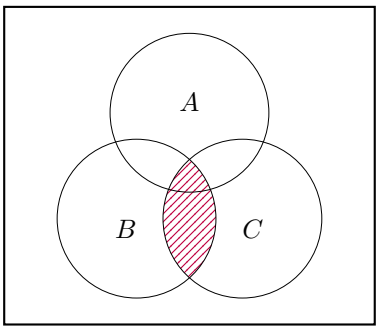
If you ever do need to prove one of these statements (or a similar one) you can use a Venn diagram or a truth table.

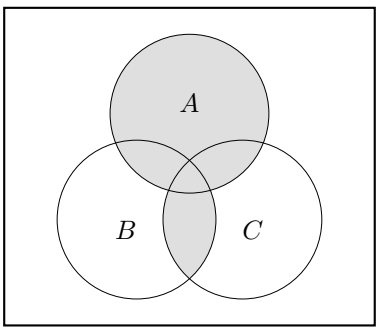
Let's prove the second distributive law,

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C),$$

with a Venn diagram. We can build the left-hand side of the law as:

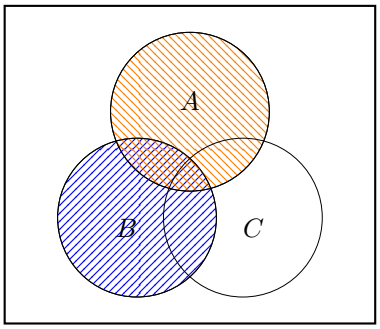


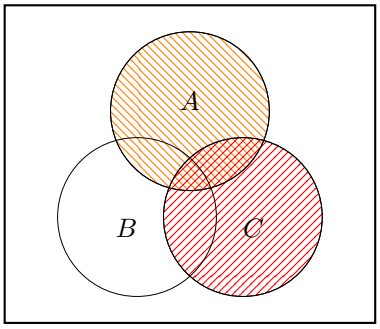


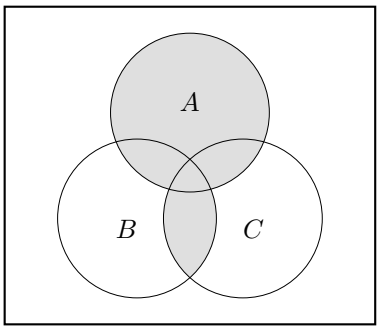


The left-hand figure is A, the middle figure is $B \cap C$, and the right-hand figure is union of these, $A \cup (B \cap C)$.

Then for the right-hand side of the law, we have:







The left-hand figure is $A \cup B$, the middle figure is $A \cup C$, and the right-hand figure is intersection of these, $(A \cup B) \cap (A \cup C)$.

We see that the areas shaded in two right-hand figures are the same, so it is indeed the case that $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.

Let's also prove the first of De Morgan's laws,

$$(A \cap B)^{\mathsf{c}} = A^{\mathsf{c}} \cup B^{\mathsf{c}},$$

this time using a truth table. (This bit might be more clear from the video above, from 14:30.) We start with a table like this, with the four possibilities of whether A and/or B are true:

A	B	$A \cap B$	$(A \cap B)^c$	A^c	B^c	$A^c \cup B^c$
False	False					
False	True					
True	False					
True	True					

We fill in the first half to find a column for the left-hand side of the law $(A \cap B)^c$:

A	B	$A \cap B$	$(A \cap B)^c$	A^c	B^c	$A^c \cup B^c$
False	False	False	**True**			
False	True	False	**True**			
True	False	False	**True**			
True	True	True	**False**			

and the second half to find a column for the right-hand side of the law $A^c \cup B^c$.

A	B	$A \cap B$	$(A \cap B)^c$	A^c	B^c	$A^c \cup B^c$
False	False	False	**True**	True	True	**True**
False	True	False	**True**	True	False	**True**
True	False	False	**True**	False	True	**True**
True	True	True	**False**	False	False	**False**

Since the $(A \cap B)^c$ column and the $A^c \cup B^c$ column are the same, these must be the same sets.

2.4 Probability axioms

Recall that, in this mathematics course, a probability will be a real number that satisfies certain properties, which we call axioms.

Definition 2.2. Let Ω be a sample space. A **probability measure** on Ω is a function $\mathbb{P}$ that assigns to each event $A \subset \Omega$ a real number $\mathbb{P}(A)$, called the **probability** of A, and that satisfies the following three axioms:

1. $\mathbb{P}(A) \geq 0$ for all events $A \subset \Omega$;
2. $\mathbb{P}(\Omega) = 1$;
3. if $A_1, A_2, \ldots$ is a finite or infinite sequence of disjoint events, then

$$\mathbb{P}(A_1 \cup A_2 \cup \cdots) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \cdots.$$

The sample space Ω together with the probability measure $\mathbb{P}$ are called a **probability space**.

Axiom 1 says that all probabilities are non-negative numbers. Axiom 2 says the probability that *something* happens is 1. Axiom 3 says that *for disjoint events* the probability that one of them happens is the sum of the individual probabilities. (Those who like their mathematical statements super-precise should note that an infinite sequence in Axiom 3 must "countable"; that is, indexed by the natural numbers $1, 2, 3. \ldots.$)

These axioms of probability (and our later results that follow from them) were first written down by the Russian mathematician Andrey Nikolaevich Kolmogorov in 1933. This marked the point from when probability theory could now be considered a proper branch of mathematics – just as legitimate as geometry or number theory – and not just a past-time that can be useful to help gamblers calculate their odds. I always find it surprising that the axioms of probability are less than 90 years old!

There are other properties that it seems natural that a probability measure should have aside from the axioms – for example, that $\mathbb{P}(A) \leq 1$ for all events A. But we will show shortly that other properties can be proven just by starting from the three axioms.

But first, let's see some examples.

Example 2.6. Suppose we wish to model tossing an biased coin the is heads with probability p, where $0 \leq p \leq 1$.

Our probability space is $\Omega = \{\mathrm{H}, \mathrm{T}\}$. The probability measure is given by

$$\mathbb{P}(\emptyset) = 0 \qquad \mathbb{P}(\{\mathrm{H}\}) = p$$
$$\mathbb{P}(\{\mathrm{T}\}) = 1 - p \qquad \mathbb{P}(\{\mathrm{H}, \mathrm{T}\}) = 1.$$

Let's check that the axioms hold:

1. Since $0 \leq p \leq 1$, all the probabilites are greater than or equal to 0.
2. It is indeed the case that $\mathbb{P}(\Omega) = \mathbb{P}(\{\mathrm{H}, \mathrm{T}\}) = 1$.
3. The only nontrivial disjoint union to check is $\{\mathrm{H}\} \cup \{\mathrm{T}\} = \{\mathrm{H}, \mathrm{T}\}$. But

$$\mathbb{P}(\{\mathrm{H}\}) + \mathbb{P}(\{\mathrm{T}\}) = p + (1 - p) = 1 = \mathbb{P}(\{\mathrm{H}, \mathrm{T}\}),$$

as required.

Example 2.7. Suppose we wish to model rolling a dice.

Our sample space is $\{1, 2, 3, 4, 5, 6\}$. The probability measure is given by

$$\mathbb{P}(A) = \frac{|A|}{6},$$

where $|A|$ is the number of sample outcomes in A.

So, for example, the probability of rolling an even number is

$$\mathbb{P}(\{2, 4, 6\}) = \frac{3}{6} = \frac{1}{2}.$$

The dice rolling is a particular case of the "classical probability" of equally likely outcomes. We'll look at this more in the next section, next week, and prove that the classical probability measure does indeed satisfy the axioms

2.5 Properties of probability

The axioms of Definition 2.2 only gave us some of the properties that we would like a probability measure to have. Our task now (in this subsection and the next) is to carefully prove how these other properties follow from just those axioms. In particular, we're not allowed to make claims that "seem likely to be true" or "are common sense" – we can only use the three axioms together with logical deductions and nothing else.

Theorem 2.1. *Let Ω be a sample space with a probability measure $\mathbb{P}$. Then we have the following:*

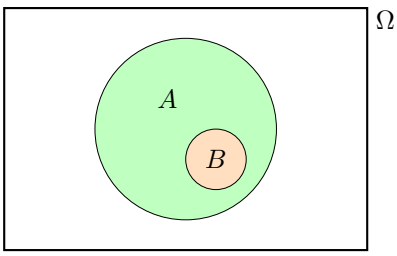
1. $\mathbb{P}(\emptyset) = 0$.
2. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ *for all events* $A \subset \Omega$.
3. *For events A and B with $B \subset A$, we have* $\mathbb{P}(B) \leq \mathbb{P}(A)$.
4. $0 \leq \mathbb{P}(A) \leq 1$ *for all events* $A \subset \Omega$.

Importantly, the third result here tells us how to deal with complements or "not" events: the probability of A *not* happening is 1 minus the probability it does happen. This is often very useful.

Proof. Statements 1 and 2 are exercise for you on Problem Sheet 2. We'll start with the third statement.

The key with most of these "prove from the axioms" problems is to think of a way to write the relevant events as part of a *disjoint* union, then use Axiom 3. Here, since B is a subset of A, it would be useful to write A as a disjoint union of B and "the bit of A that isn't in B. That is, we have the disjoint union

$$B \cup (A \cap B^c) = A.$$



Applying Axiom 3 to this disjoint union gives

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

We're happy to see the first term on the left-hand side and the term on the right-hand side. But what about the awkward $\mathbb{P}(A \cap B^{\mathsf{c}})$? Well, by Axiom 1, we know that $\mathbb{P}(A \cap B^{\mathsf{c}}) \geq 0$, and hence

$$\mathbb{P}(B) + 0 \leq \mathbb{P}(A),$$

and we are done with the third statement.

For the fourth statement, we have $\mathbb{P}(A) \geq 0$ directly from Axiom 1, so only need to show that $\mathbb{P}(A) \leq 1$. We can do this using the third statement of this theorem. For any event $A \subset \Omega$, the third statement tells us that $\mathbb{P}(A) \leq \mathbb{P}(\Omega)$. But Axiom 2 tells us that $\mathbb{P}(\Omega) = 1$, so we are done. □

2.6 Addition rules for unions

If we have two or more events, we'd like to work out the probability of their union; that is, the probability that at least one of them occurs.

We already have an addition rule for *disjoint* unions.

Theorem 2.2. *Let $A, B \subset \Omega$ be two disjoint events. Then*

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

Proof. In Axiom 3, take the finite sequence $A_1 = A$, $A_2 = B$. □

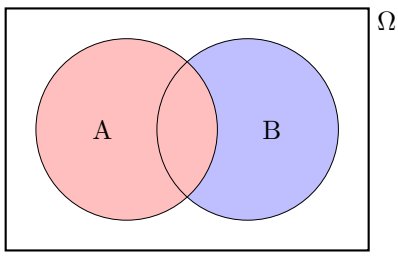
But what about if A and B are not disjoint? Then we have the following.

Theorem 2.3. *Let $A, B \subset \Omega$ be two events. Then*

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

You may have seen this result before. You've perhaps justified it by saying something like this: "We can add the two probabilities together, except now we've double-counted the overlap, so we have to take the probability of that away." Maybe you drew a Venn diagram. That's OK as a way to remember the result – but this is a proper university mathematics course, so we have to carefully *prove* it starting from just the axioms and nothing else.

As always, the key is to find a way of writing $A \cup B$ as a *disjoint* union. (In general, $A \cup B$ can be a non-disjoint union that has an overlap.) Well, if we want $A \cup B = A \cup \{\text{something}\}$ to be a *disjoint* union, then the "something" will have to be the bit of B that's not also in A, which is $B \cap A^{\mathsf{c}}$.



Proof. First note, following the discussion above, that we have

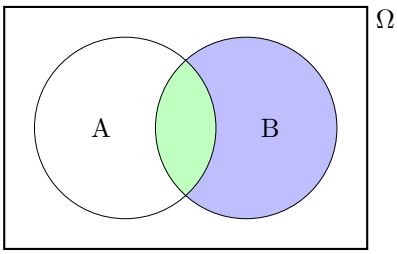
$$A \cup B = A \cup (B \cap A^{\mathsf{c}}),$$

where the union on the right is of the disjoint events A and $B \cap A^{\mathsf{c}}$. Therefore we can use Axiom 3 to get

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B \cap A^{\mathsf{c}}). \tag{2.1}$$

The left-hand side looks good, and the first term on the right-hand side looks good. To deal with the second term on the right-hand side, we need to write it down as part of a disjoint union again. Can we find another one? Yes! We can use $B \cap A^{\mathsf{c}}$ together with $B \cap A$ to build the whole of B. So have a disjoint union

$$(B \cap A^{\mathsf{c}}) \cup (B \cap A) = B.$$



Since this union is disjoint, we can use Axiom 3 again, to get

$$\mathbb{P}(B \cap A^{\mathsf{c}}) + \mathbb{P}(B \cap A) = \mathbb{P}(B).$$

Rearranging this gives

$$\mathbb{P}(B \cap A^{\mathsf{c}}) = \mathbb{P}(B) - \mathbb{P}(B \cap A). \tag{2.2}$$

Finally, substituting (2.2) into (2.1) gives

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

as required. □

Example 2.8. *Consider picking a card from a deck at random, with* $\mathbb{P}(A) = |A|/52$. *What's the probability the card is a spade or an ace?*

It is possible to just to work this out directly. But let's use our addition law for unions.

We have $\mathbb{P}(\text{spade}) = \frac{13}{52}$ and $\mathbb{P}(\text{ace}) = \frac{4}{52}$. So we have

$$\mathbb{P}(\text{spade or ace}) = \tfrac{13}{52} + \tfrac{4}{52} - \mathbb{P}(\text{spade and ace}).$$

But $\mathbb{P}(\text{spade and ace})$ is the probability of picking the ace of spades, which is $\frac{1}{52}$. Therefore

$$\mathbb{P}(\text{spade or ace}) = \tfrac{13}{52} + \tfrac{4}{52} - \tfrac{1}{52} = \tfrac{16}{52} = \tfrac{4}{13}.$$

Similar addition rules can be proven in the same way for unions of more events. For three events, we have

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

Note that we add the probabilities of individual events, then subtract the probabilities of pairs, then add the probability of the triple.

The **inclusion–exclusion principle** is the general rule:

$$\begin{aligned}\mathbb{P}(A_1 \cup A_2 \cup \cdots \cup A_n) &= \sum_i \mathbb{P}(A_i) - \sum_{i \neq j} \mathbb{P}(A_i \cap A_j) \\ &\quad + \sum_{i \neq j \neq k} \mathbb{P}(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n-1} \mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n),\end{aligned}$$

where we continue by subtracting the probabilities of quadruples, adding the probabilities of five events, etc.

Summary

- A sample space Ω is a set representing all possible sample outcomes. An event is a subset of Ω.
- For events A and B, we also have the complement "not A" A^{c}, the intersection "A and B" $A \cap B$, and the union "A or B" $A \cup B$.
- The axioms of probability are (1) $\mathbb{P}(A) \geq 0$; (2) $\mathbb{P}(\Omega) = 1$; and (3) that for disjoint events $A_1, A_2, \ldots$, we have $\mathbb{P}(A_1 \cup A_2 \cup \cdots) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \cdots$.
- Other properties can be proven from these axioms, like the complement rule $\mathbb{P}(A^{\mathsf{c}}) = 1 - \mathbb{P}(A)$, and the addition rule for unions $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

Chapter 3

Classical probability

3.1 Probability with equally likely outcomes

Classical probability is the name we give to probability where there are a finite number of equally likely outcomes.

Classical probability was the first type of probability to be formally studied – partly because it is the simplest, and partly because it was useful for working out how to win at gambling. Tossing fair coins, rolling dice, and dealing cards are all common gambling situations that can be studied using classical probability – in a deck of cards, for example, there are 52 cards that are equally likely to be drawn. Among the first works to seriously study classical probability were "Book on Games of Chance" by Girolamo Cardano (written in 1564, but not published until 1663, one hundred years later), and a famous series of letters letters between Blaise Pascal and Pierre de Fermat in 1654.

Definition 3.1. Let Ω be a finite sample space. Then the **classical probability measure** on Ω is given by

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

So to work out a classical probability $\mathbb{P}(A)$, crucially we need to be able to count how many outcomes $|A|$ are in the event A and count how many outcomes $|\Omega|$ are in the whole sample space Ω. (This is why classical probability is also called "enumerative probability" – "enumeration" is another word for counting.) In this section, we'll look at some different ways in which we can count the number of outcomes in common events and sample spaces.

There's something we ought to check before going any further!

Theorem 3.1. *Let Ω be a finite nonempty sample space. Then the classical probability measure on Ω,*

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

is indeed a probability measure, in that is satisfies the three axioms in Definition 2.2.

Proof. We'll take the axioms one by one.

1. Since $|\Omega| \geq 1$ and $|A| \geq 0$, it is indeed the case that $\mathbb{P}(A) = |A|/|\Omega| \geq 0$.
2. We have $\mathbb{P}(\Omega) = \dfrac{|\Omega|}{|\Omega|} = 1$, as required.
3. Since we have a finite sample space, we only need to show Axiom 3 for a sequence of two disjoint events; the argument can be repeated to get any finite number of events. Let $A = \{a_1, a_2, \dots, a_k\}$ and $B = \{b_1, b_2, \dots, b_l\}$ be two disjoint events with $|A| = k$ and $|B| = l$. Note that we can enumerate the elements of the disjoint union $C = A \cup B$ as

$$c_1 = a_1, c_2 = a_2, \dots, c_k = a_k, c_{k+1} = b_1, c_{k+2} = b_2, \dots, c_{k+l} = b_l.$$

   Since A and B are disjoint, this list has no repeats, and we see that $|C| = |A \cup B| = k + l$. Hence

$$\mathbb{P}(A \cup B) = \frac{k+l}{|\Omega|} = \frac{k}{|\Omega|} + \frac{l}{|\Omega|} = \mathbb{P}(A) + \mathbb{P}(B),$$

   and Axiom 3 is fulfilled.

□

3.2 Multiplication principle

In classical probability, to find the probability of an event A, we need to count the number of outcomes in A and the total number of possible outcomes in Ω. This can be easy when we're just looking at one choice – like the 2 outcomes from tossing a single coin, the 6 outcomes of rolling a single dice, or the 52 outcomes from dealing a single card. Now we're going to look at what happens if there are a number of choices one after another – like tossing multiple coins, rolling more than one dice, or dealing a hand of cards.

Here, an important principle is the **multiplication principle**. The multiplication principle says that if you have n choices followed by m choices, than all together you have $n \times m$ total choices. You can see this by imagining the choices in a $n \times m$ grid, with the n columns representing the first choice and m rows representing the second choice. For example, suppose you go to a burger restaurant where there are 3 choices of burger (beefburger, chicken burger, veggie burger) and 2 choices of sides (fries, salad), then altogether there are $3 \times 2 = 6$ choices of meal.

	Beefburger	Chicken burger	Veggie burger
Fries	1: Beefburger with fries	2: Chicken burger with fries	3: Veggie burger with fries

	Beefburger	Chicken burger	Veggie burger
Salad	4: Beefburger with salad	5: Chicken burger with salad	6: Veggie burger with salad

More generally, if you have m stages of choosing, with n_1 choices in the first stage, then n_2 choices in the second stage, all the way to n_m choices in the final stage, you have $n_1 \times n_2 \times \cdots \times n_m$ total choices altogether.

Example 3.1. *Five fair coins are tossed. What is the probability they all show the same face?*

Here, the sample space Ω is the set of all sequences of 5 coin outcomes. How many sample outcomes are in Ω? Well, the first coin can be heads or tails (2 choices); the second coin can be heads or tails (2 choices) and so on, until the fifth and final coin. So, by the multiplication principle, $|\Omega| = 2 \times 2 \times 2 \times 2 \times 2 = 2^5 = 32$.

The event we're interested in is $A = \{\text{HHHHH}, \text{TTTTT}\}$, the event that the faces are all the same – either all heads or all tails. This clearly has $|A| = 2$ outcomes.

So the probability all five coins show the same face is

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{2}{32} = \frac{1}{16} \approx 0.06.$$

Example 3.2. *Five dice are rolled. What is the probability we get at least one 6?*

Here, Ω is the set of all possible sequences of 5 dice rolls. Clearly $|\Omega| = 6^5 = 7776$.

Also, A is the set of all dice roll sequences with at least one 6. Whenever you see a question with the phrase "at least one" in it, it's very often to look at the complementary event A^c instead. We know from the last section that $\mathbb{P}(A) = 1 - \mathbb{P}(A^c)$, but in "at least one" questions, it's often easier to count $|A^c|$ than to count $|A|$.

Here, is A is the set of all dice roll sequences with at least one 6, then A^c is the set of dice roll sequence with no 6 at all. This means all five dice must have rolled a 1, 2, 3, 4, or 5. Since each of the five dice rolls has 5 possibilities, this means that $|A^c| = 5^5 = 3125$.

Finally, we see that

$$\mathbb{P}(A) = 1 - \mathbb{P}(A^c) = 1 - \frac{|A^c|}{|\Omega|} = 1 - \frac{5^5}{6^5} = \frac{4651}{7776} \approx 0.70.$$

3.3 Sampling with and without replacement

Example 3.3. *A bag contains 15 balls: 10 black balls and 5 white balls. We draw 3 balls out of the bag. What is the probability all 3 balls are black* ***(a)*** *if*

we put each ball back into the bag after it is chosen; ***(b)*** *if we do not put each ball back into the bag after it is chosen.*

Let's start with (a). The number of ways to choose a ball out 15 on three occasions is $|\Omega| = 15^3$. The number of ways to choose a black balls out of 10 on three occasions is $|A| = 10^3$. Hence

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{10^3}{15^3} = \frac{1000}{3375} = \frac{8}{27} \approx 0.30.$$

What about (b)? Here we don't put the ball back in the bag once it has been chosen. There are 15 ways to pick the first ball. But then there are only 14 balls left in the bag for the second choice, and only 13 balls for the third choice. So $|\Omega| = 15 \times 14 \times 13$. Similarly, there are 10 ways the first ball can be black. But once that black ball is removed, only 9 choices for the second black ball, and only 8 for the third. So $|A| = 10 \times 9 \times 8$. So this time we have

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{10 \times 9 \times 8}{15 \times 14 \times 13} = \frac{720}{2730} = \frac{24}{91} \approx 0.26,$$

which is smaller than the answer in part (a).

This example illustrated the difference between **sampling with replacement** (when the balls were put back into the bag) and **sampling without replacement** (when the balls were not put back). If we want to sample k items from a set of n items, then:

- the number of ways to sample with replacement is

$$n^k = n \times n \times \cdots \times n;$$

- the number of ways to sample without replacement is

$$n^{\underline{k}} = n \times (n-1) \times \cdots \times (n-k+1).$$

Here, we've defined the notation $n^{\underline{k}}$ for the number of ways to sample without replacement; this is called the **falling factorial** or **permutation number**. (Notice that the subscript is underlined here; other notations include $(n)_k$, $P(n,k)$, or ${}^nP_k$.)

3.4 Ordering

Example 3.4. *Suppose a lecturer marks a pile of n exam papers, all of which receive a different mark. What is the probability she ends up marking them in order from lowest scoring first in the pile to highest scoring last in the pile?*

Here, the sample space Ω is the set of all orderings of the n exam papers by mark, and A is the event that the papers are in order from lowest to highest scoring. It's clear that $|A| = 1$: since the exams scored different marks, there's

only one way of putting the exams in the correct lowest-to-highest order. But what's $|\Omega|$?

There are n choices for the first exam paper to be marked. Then, for the second exam paper, there are $n-1$ choices left, because I'm not going to mark the same paper twice. There are $n-2$ choices for the third exam paper. And so on, until I have marked $n-1$ papers, and there is only 1 choice left for the final paper. So we have

$$|\Omega| = n^{\underline{n}} = n(n-1)(n-2)\cdots 3\cdot 2\cdot 1 = n!$$

ways to order the exam papers.

Hence, the probability the papers are marked in order is

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{1}{n(n-1)\cdots 2\cdot 1} = \frac{1}{n!}.$$

This number

$$n! = n^{\underline{n}} = n(n-1)(n-2)\cdots 3\cdot 2\cdot 1$$

is called n **factorial** and denoted $n!$. It is the number of ways that n different objects can be ordered.

It can sometimes be useful to write the falling factorial $n^{\underline{k}}$ in terms of the factorial, like this:

$$\begin{aligned} n^{\underline{k}} &= n(n-1)\cdots(n-k+1) \\ &= n(n-1)\cdots(n-k+1)\times\frac{(n-k)(n-k-1)\cdots 2\cdot 1}{(n-k)(n-k-1)\cdots 2\cdot 1} \\ &= \frac{n(n-1)\cdots(n-k+1)(n-k)(n-k-1)\cdots 2\cdot 1}{(n-k)(n-k-1)\cdots 2\cdot 1} \\ &= \frac{n!}{(n-k)!}. \end{aligned}$$

Example 3.5. Suppose you shuffle a pack of cards. The resulting ordering of the deck has 52! possibilities. This is an unimaginably huge number – it's roughly

$$52! \approx 8\times 10^{67};$$

that is, an 8 followed by 67 zeroes.

In comparison, the universe has existed for about 4×10^{17} seconds, and there are about 7×10^{9} people alive. If every person on the planet had shuffled a deck of cards one million times a second for the entire lifetime of the universe, they could only expect to have got through

$$(4\times 10^{17})\times(7\times 10^{9})\times 10^{6} \approx 3\times 10^{33}.$$

This is only the most tiny, microscopic fraction of 52!. So every time you have ever shuffled a deck of cards, it is essentially certain that you have created an ordering of the deck that has never existed before.

3.5 Sampling without replacement in any order

Example 3.6. *In "Lotto", the UK national lottery, you can buy a ticket for £2 and choose 6 numbers between 1 and 59. If your 6 numbers match the 6 numbers chosen by the lottery machine, you win the jackpot (usually between £2 million and £20 million, shared between the tickets that get all 6 numbers). If you buy a ticket, what is the probability you win the jackpot?*

Here, Ω is the set of all possible sets of 6 winning numbers, and A is the set of numbers on your ticket. Clearly $|A| = 1$, but what is $|\Omega|$?

Well, the first number out of the machine has 59 possibilities, the second number has 58 possibilities, and so on, making

$$59 \times 58 \times 57 \times 56 \times 55 \times 54 = 59^{\underline{6}}.$$

But this isn't the correct answer, because the same set of numbers could be drawn from the machine in any order! The sets of numbers $\{1, 2, 3, 4, 5, 6\}$ and $\{1, 2, 3, 4, 6, 5\}$ and $\{6, 5, 4, 3, 2, 1\}$ are all the same set of numbers. How many ways can we see the same list of numbers? This is precisely the number of orderings of 6 numbers, which we know is 6!. So the number of possible sets of 6 numbers to come out of the machine is actually

$$\binom{59}{6} = \frac{59^{\underline{6}}}{6!} = \frac{59 \times 58 \times 57 \times 56 \times 55 \times 54}{6 \times 5 \times 4 \times 3 \times 2 \times 1} \approx 45 \text{ million}.$$

Thus the probability that your ticket wins the jackpot is

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{1}{\binom{59}{6}} \approx \frac{1}{45 \text{ million}} \approx 0.00000002.$$

Here, we have introduced the notation

$$\binom{n}{k} = \frac{n^{\underline{k}}}{k!} = \frac{n(n-1)\cdots(n-k+1)}{k(k-1)\cdots 2 \cdot 1}$$

for the number of ways to choose k objects out of n without replacement *and where the order they were chosen in doesn't matter.* This is called the **binomial coefficient**, although when we say it out loud we normally just say **"n choose k"**. (Another notation for the binomial coefficient is ${}^nC_k$.)

It can sometimes be useful to remember that $n^{\underline{k}} = n!/(n-k)!$ allows us to write the binomial coefficient in terms of the factorial function as

$$\binom{n}{k} = \frac{n^{\underline{k}}}{(n-k)!} = \frac{n!}{k!(n-k)!}.$$

Example 3.7. *You are dealt a "hand" of 13 cards from a deck of 52 cards. What is the probability that you have the Ace, King, Queen, and Jack of Spades?*

Here, Ω is the set of all 13-card hands from the deck, and A is the subset of those that contain the AKQJ of Spades.

Using the binomial coefficient notation, it's clear that

$$|\Omega| = \binom{52}{13} = \frac{52 \times 51 \times \cdots \times 41 \times 40}{13 \times 12 \times \cdots \times 2 \times 1}.$$

What about $|A|$? If we fix the fact that the hand contains the 4 cards AKQJ of Spades, then it also contains $13 - 4 = 9$ cards out of the other $52 - 4 = 48$ remaining cards in the deck. This makes

$$|A| = \binom{48}{9} = \frac{48 \times 47 \times \cdots \times 41 \times 40}{9 \times 8 \times \cdots \times 2 \times 1}$$

hands.

Thus the probability that the hand contains AKQJ of Spades is

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{\binom{48}{9}}{\binom{52}{13}}.$$

Conveniently, we can simplify the expression quite a lot, because plenty of cancellation will go on. We have

$$\begin{aligned}
\mathbb{P}(A) = \frac{\binom{48}{9}}{\binom{52}{13}} &= \frac{\frac{48\times47\times\cdots\times41\times40}{9\times8\times\cdots\times2\times1}}{\frac{52\times51\times\cdots\times41\times40}{13\times12\times\cdots\times2\times1}} \\
&= \frac{48 \times 47 \times \cdots \times 41 \times 40}{52 \times 51 \times \cdots \times 41 \times 40} \times \frac{13 \times 12 \times \cdots \times 2 \times 1}{9 \times 8 \times \cdots \times 2 \times 1} \\
&= \frac{13 \times 12 \times 11 \times 10}{52 \times 51 \times 50 \times 49} \\
&\approx 0.0026,
\end{aligned}$$

or about 1 in every 380 hands.

There's one other useful fact about the binomial coefficient we should mention.

Theorem 3.2.

$$\binom{n}{k} = \binom{n}{n-k}.$$

We'll give two different proofs of this fact.

Proof. We can use the formula for the binomial coefficient in terms of the factorial function. This formula gives

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \qquad \binom{n}{n-k} = \frac{n!}{(n-k)!k!}.$$

It's clear from the factorial expression that these two quantities are equal. □

Proof. Suppose we have n balls, and we want to paint k of the red and the other $n - k$ blue. How many ways can we do this?

One way is to say there are $\binom{n}{k}$ ways to choose the k balls to paint red, then we are forced paint the other $n-k$ blue.

Another way is to say there are $\binom{n}{n-k}$ to choose the $n-k$ balls to paint blue, then we are forced to paint the other k red.

We have "double counted" the same quantity two different ways, so the answer must be the same, so

$$\binom{n}{k} = \binom{n}{n-k}.$$

□

Which proof do you prefer? (I strongly prefer the second proof, because I think it doesn't just *verify* that the theorem is true, but further explains *why* the result is true.)

3.6 Birthday problem

There are $k = 23$ students in a class. What is the probability that at least two of the students share a birthday?

This a famous problem, known as the "birthday problem". You may have seen this problem before. But let's try to solve it using the techniques from this section of notes. (We'll assume all days are equally likely for birthdays, and ignore the effect of leap days.)

The sample space $|\Omega|$ is the set of possible birthdays for the k students. Clearly $|\Omega| = 365^k$.

Let A be the even that at least two students share a birthday. Since this is an "at least" event, it seems like it might be a good idea to look instead at the complement event A^{c} instead. If A is the event that there's at least one shared birthday, then A^{c} is the event that there are *no* shared birthdays; that is, A^{c} is the event that all k students have *different* birthdays.

So what is $|A^{\mathsf{c}}|$, the number of ways the k students can have different birthdays? Well, the first student can have any of the 365 days for their birthday. For them to have different birthdays, the second student only has 364 days available. Then the third student must avoid the birthday of students 1 and 2, so has 363 available days, and so on. We see that

$$|A^{\mathsf{c}}| = 365 \times 364 \times \cdots \times (365 - k + 1) = 365^{\underline{k}}.$$

Hence, the probability at least two students share a birthday is

$$\mathbb{P}(A) = 1 - \mathbb{P}(A^{\mathsf{c}}) = 1 - \frac{365^{\underline{k}}}{365^k} = 1 - \frac{365}{365} \cdot \frac{364}{365} \cdots \frac{365-k+1}{365}.$$

Setting $k = 23$, we can calculate the required answer in R:

```
k <- 23
1 - prod((365:(365 - k + 1)) / 365)
```

```
## [1] 0.5072972
```

The probability is 50.7%. So it's more likely than not that at least two students share a birthday.

Some people find it surprising that only 23 students have such a high probability of sharing a birthday, since 23 is so small compared to 365. But remember there are $\binom{23}{2} = 253$ *pairs* of birthdays, each of which is a potential match, and 253 is quite a big number.

Summary

- "Classical probability" describes the situation where there are finitely many equally likely outcomes. The classical probability $\mathbb{P}(A) = |A|/|\Omega|$ requires us to count how many outcomes there are in events or sample spaces.
- The multiplication principle says that n choices followed by m choices makes $n \times m$ choices in total.
- Sampling k objects out of n with replacement gives n^k choices.
- Sampling k objects out of n without replacement gives $n^{\underline{k}} = n(n-1) \dots (n-k+1)$ choices.
- Ordering n objects can be done in $n! = n^{\underline{n}} = n(n-1) \dots 2 \cdot 1$ ways.
- The number of ways to sample k objects out of n when the order doesn't matter is given by the binomial coefficient $\binom{n}{k} = n^{\underline{k}}/k!$.

Problem Sheet 2

This is Problem Sheet 2. This problem sheet covers Section 2 and Section 3 of the notes. You should work through all the questions on this problem sheet during Weeks 2 and 3, in preparation for your tutorial in Week 4. The problem sheet contains two assessed questions, which are due in by **2pm on Monday 25 October**.

A: Short questions

A1. Suppose you toss a coin 10 times. What would you suggest for a sample space **(a)** if you only care about the total number of heads; **(b)** if you care about the result of every coin toss?

A2. Let A, B and C be events in a sample space Ω. Write the following events using only A, B, C and the complement, intersection, and union operations.

(a) C happens but A doesn't.

(b) At least one of A, B and C happens.

(c) Exactly one of B or C happens.

(d) Exactly two of A, B and C happens.

A3. Let Ω be a sample space with a probability measure $\mathbb{P}$, and let $A, B \subset \Omega$ be events. State, with brief explanations, whether the following statements are true or false:

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

(b) $\mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^{\mathsf{c}}) = \mathbb{P}(A)$.

(c) $\mathbb{P}(A \cup B) \leq \mathbb{P}(A)$

(d) If A and B are disjoint, then $\mathbb{P}((A \cup B)^{\mathsf{c}}) = 1 - \mathbb{P}(A) - \mathbb{P}(B)$.

A4. What is the value of the following expressions?

(a) $6!$

(b) $8^{\underline{4}}$

(c) $\binom{10}{4}$

A5. An urn contains 5 red balls and 7 blue balls. Four balls are drawn from the urn. What is the probability that at least one of the balls is red, if the balls are drawn **(a)** with replacement; **(b)** without replacement?

B: Long questions

B1 Starting from just the three probability axioms, prove the following statements:

(a) $\mathbb{P}(\emptyset) = 0$.

(b) $\mathbb{P}(A^{\mathsf{c}}) = 1 - \mathbb{P}(A)$.

B2. Suppose we pick a number at random from the set $\{1, 2, \dots, 2021\}$.

(a) What is the probability that the number is divisible by 5?

(b) What is the probability the number is divisible by 5 or by 7?

B3. In this question, you will have to use the standard two-event form of the addition rule for unions

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

(a) Using the two-event addition rule, show that

$$\mathbb{P}(C \cup D \cup E) = \mathbb{P}(C) + \mathbb{P}(D \cup E) - \mathbb{P}(C \cap (D \cup E)).$$

(b) Using the two-event addition rule and the distributive law, or otherwise, prove the three-event form of the addition rule for unions:

$$\mathbb{P}(C \cup D \cup E) = \mathbb{P}(C) + \mathbb{P}(D) + \mathbb{P}(E) - \mathbb{P}(C \cap D) - \mathbb{P}(C \cap E) - \mathbb{P}(D \cap E) + \mathbb{P}(C \cap D \cap E).$$

B4. Eight friends are about to sit down at random at a round table. Find the probability that

(a) Ashley and Brook sit next to each other, with Chris directly opposite Brook;

(b) neither Ashley, Brook nor Chris sit next to each other.

B5. Suppose your tutorial group contains 12 students – you and 11 others. The tutor wishes to choose 4 members of the group to present their work.

(a) How many ways can the tutor choose the presentation group?

(b) How many ways can the tutor choose the presentation group if you are one of the presenters?

(c) How many ways can the tutor choose the presentation group if you are *not* one of the presenters?

(d) Pascal's formula says that

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

Prove Pascal's formula.

C: Assessed questions

The last two questions are **assessed questions**. These two questions count for 3% of your final mark for this module.

The deadline for submitting your solutions is **2pm on Monday 25 October** at the beginning of Week 5, although I strongly recommend completing and submitting your work during Week 4. Submission will be via Gradescope; submission will open on Monday 18 October. Your work will be marked by your tutor and returned on Monday 1 November, when solutions will also be made available.

Both questions are "long questions", where the marks are not only for mathematical accuracy but also for the clarity and completeness of your explanations.

You should not collaborate with others on the assessed questions: your answers must represent solely your own work. The University's rules on academic integrity – and the related punishments for violating them – apply to your work on the assessed questions.

C1. Let A and B be two events with $\mathbb{P}(A) = 0.8$ and $\mathbb{P}(B) = 0.4$. The following questions concern the value of $\mathbb{P}(A \cap B)$, the probability that both A and B occur.

(a) Prove the upper bound $\mathbb{P}(A \cap B) \leq 0.4$.

(b) Prove that this upper bound can be achieved, by giving an example of a sample space Ω, a probability measure $\mathbb{P}$ and events $A, B \subset \Omega$ such that $\mathbb{P}(A) = 0.8$ and $\mathbb{P}(B) = 0.4$, with equality $\mathbb{P}(A \cap B) = 0.4$ in the upper bound.

(c) Give a lower bound for $\mathbb{P}(A \cap B)$ – that is, prove that $\mathbb{P}(A \cap B) \geq$ something – and show that this lower bound can be achieved. (Try to work out the correct bound, even if you can't formally prove it.)

C2. A "random digit" is a number chosen at random from $\{0, 1, \dots, 9\}$, each with equal probability. A statistician choses n random digits.

(a) For $k = 0, 1, \dots, 9$, let A_k be the event that all the digits are k or smaller. What is the probability of A_k, as a function of k and n?

(b) Let B_k be the event that the largest digit chosen is equal to k. What is the probability of B_k? Justify your answer carefully.

Solutions to short questions

A1. (a) $\{0, 1, \dots, 10\}$ (b) The set of all 2^{10} vectors of length 10 with entries from $\{\mathrm{H}, \mathrm{T}\}$. **A2.** (a) $C \cap A^{\mathsf{c}}$ (b) $A \cup B \cup C$ (c) $(B \cup C) \cap (B \cap C)^{\mathsf{c}}$ or $(B \cap C^{\mathsf{c}}) \cup (B^{\mathsf{c}} \cap C)$ (d) $(A \cap B \cap C^{\mathsf{c}}) \cup (A \cap B^{\mathsf{c}} \cap C) \cup (A^{\mathsf{c}} \cap B \cap C)$ or other equivalent. **A3.** (a) False (b) True (c) False (d) True **A4.** (a) 720 (b) 1680 (c) 210 **A5.** (a) 0.884 (b) 0.929

Chapter 4

Independence and conditional probability

4.1 Independent events

Suppose 40% of people have blond hair, and 20% of people have blue eyes. What proportion of people have both blond hair and blue eyes?

The answer to this question is: we don't know. The question doesn't give us enough information to tell. However, *if* it were the case that having blond hair didn't effect your chance of having blue eyes, then we could work out the answer. If that were true, we would think that the 20% of people with blue eyes equally made up both 20% of the blonds and also 20% of the non-blonds. Thus the proportion of people with blond hair and blue eyes would be this 20% of the 40% of people with blond hair, and 20% of 40% is $0.2 \times 0.4 = 0.08$, or 8%.

To put it in probability language, *if* blond hair and blue eyes were unrelated, then we would expect that

$$\mathbb{P}(\text{blond hair and blue eyes}) = \mathbb{P}(\text{blond hair}) \times \mathbb{P}(\text{blue eyes}).$$

This is an important property known as "independence".

Definition 4.1. Two events A and B are said to be **independent** if

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

There are two ways we can use this definition.

- If we know $\mathbb{P}(A)$, $\mathbb{P}(B)$, and $\mathbb{P}(A \cap B)$, then we can find out whether or not A and B are independent by checking whether or not $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$.
- If we know $\mathbb{P}(A)$ and $\mathbb{P}(B)$ and we know that A and B are independent, then we can find $\mathbb{P}(A \cap B)$ by calculating $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$.

In this second case, we might know A and B are independent because we are specifically told they are. But we might reason that A and B because the related experiments are not physically related – for example if we roll a dice then toss a coin, we might reason that {roll a 5} and {the coin lands Heads} must be independent because the dice roll doesn't effect the coin toss, and use the independence assumption in calculations.

Example 4.1. *Consider rolling a dice. Let $A = \{\text{even number}\} = \{2, 4, 6\}$, and let $B = \{\text{roll at least 4}\} = \{4, 5, 6\}$. Are A and B independent?*

Clearly we have $\mathbb{P}(A) = \frac{3}{6} = \frac{1}{2}$ and $\mathbb{P}(B) = \frac{3}{6} = \frac{1}{2}$. The intersection is $A \cap B = \{4, 6\}$, so $\mathbb{P}(A \cap B) = \frac{2}{6} = \frac{1}{3}$. So we see that

$$\mathbb{P}(A \cap B) = \frac{1}{3} \qquad \text{and} \qquad \mathbb{P}(A)\,\mathbb{P}(B) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}.$$

So $\mathbb{P}(A \cap B) \neq \mathbb{P}(A)\,\mathbb{P}(B)$, and the two events are not independent.

Example 4.2. *A biased coin has probability p of landing Heads and probability $1 - p$ of landing Tails. You toss the coin 3 times. Assuming tosses of the coin are independent, calculate the probability of getting exactly 2 Heads.*

There are three ways we could get exactly 2 Heads: HHT, HTH, or THH. For the first of these,

$$\mathbb{P}(\text{HHT}) = \mathbb{P}(\text{first coin H} \cap \text{second coin H} \cap \text{third coin T}).$$

Since tosses of the coin are independent, we therefore have

$$\begin{aligned}\mathbb{P}(\text{HHT}) &= \mathbb{P}(\text{first coin H}) \times \mathbb{P}(\text{second coin H}) \times \mathbb{P}(\text{third coin T}) \\ &= p \times p \times (1 - p) \\ &= p^2(1 - p).\end{aligned}$$

Similarly,

$$\mathbb{P}(\text{HTH}) = \mathbb{P}(\text{THH}) = p^2(1 - p)$$

also.

Finally, because the events are disjoint, we have

$$\mathbb{P}(\text{HHT} \cup \text{HTH} \cup \text{THH}) = \mathbb{P}(\text{HHT}) + \mathbb{P}(\text{HTH}) + \mathbb{P}(\text{THH}) = 3p^2(1 - p).$$

4.2 Conditional probability

Let us to return to the example of blond hair and blue eyes. Suppose the population statistics are like this:

	Brown hair	**Blond hair**	**Total**
Brown eyes	50%	30%	80%
Blue eyes	10%	10%	20%

	Brown hair	**Blond hair**	**Total**
Total	60%	40%	100%

(It turns out that $\mathbb{P}(\text{blond hair and blue eyes}) = 0.1 \neq 0.08$, so they are not independent.)

We know that 20% of people have blue eyes. But suppose you already know that someone has blond hair: what then is their probability of have blue eyes *given* that they have blond hair?

Well, the 40% of blond-haired people is made up of the 10% of people who also have blue eyes to go along with their blond hair, and the 30% of people who have brown eyes to go along with their blond hair. So of the 40% of blond-haired people, three times as many have brown eyes, so only one quarter of that 40% have blue eyes. If we use a vertical line | in a probability to mean "given" (or "assuming that" or "conditional upon"), then we can write this as

$$\mathbb{P}(\text{blue eyes} \mid \text{blond hair}) = \frac{\mathbb{P}(\text{blue eyes and blond hair})}{\mathbb{P}(\text{blond hair})} = \frac{0.1}{0.4} = \frac{1}{4}.$$

What we've seen here is called a "conditional probability".

Definition 4.2. Let A and B be events, with $\mathbb{P}(A) > 0$. Then the **conditional probability of B given A** is defined to be

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

The condition $\mathbb{P}(A) > 0$ is to ensure we don't have any "divide by 0" errors. (I normally won't bother saying this explicitly – any statement about conditional probability will implicitly assume that the event being conditioned on has nonzero probability.)

As with independence, conditional probability can be used in different ways: given any two of $\mathbb{P}(A)$, $\mathbb{P}(A \cap B)$, and $\mathbb{P}(B \mid A)$ you can work out the other one.

Conditional probability ties in with independence in an important way. Suppose A and B are independent, so $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$. Then the conditional probability becomes

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

so $\mathbb{P}(B \mid A) = \mathbb{P}(B)$. In other words, if A and B are independent, then A happening doesn't affect the probability of B happening (and vice versa).

So when we have independence, $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$, and the mathematics is quite easy. But conditional probability tells us how things work when we don't have independence.

4.3 Chain rule

We can rewrite the definition of conditional probability like this:

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

This can be a useful way to think when A concerns the first stage of an experiment and B the second stage. This says that the probability A happens then B happens is equal to the probability A happens multiplied the probability, given that A has happened, that B then happens too.

We can extend this to more events. For three events, we have

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

which can be useful when we have three stages of an experiment.

Continuing that process, we get a general rule.

Theorem 4.1 (Chain rule)**.** *For events $A_1, A_2, \dots, A_n$, we have*

$$\begin{aligned}&\mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n)\\ &\qquad = \mathbb{P}(A_1)\,\mathbb{P}(A_2 \mid A_1)\,\mathbb{P}(A_3 \mid A_1 \cap A_2) \cdots \mathbb{P}(A_n \mid A_1 \cap A_2 \cap \cdots \cap A_{n-1}).\end{aligned}$$

Often questions that can be solved using the classical probability counting methods from Section 3 also be solved in stages using the chain rule. (It's a matter of personal taste which you prefer.)

Example 4.3. *Recall the Lotto problem from Example 3.6: What is the probability we match 6 balls from 59?*

Let $A_1, A_2, \dots, A_6$ be the events that the first, second, ..., sixth balls out of the machine are on our ticket. Clearly $\mathbb{P}(A_1) = \frac{6}{59}$, as we have six numbers the ball could match. Then the conditional probability that the second ball matches given that the first ball matched is $\mathbb{P}(A_2 \mid A_1) = \frac{5}{58}$, because there are 58 balls left in the machine and, given that we got the first number right, there are 5 numbers left on our ticket. Similarly, $\mathbb{P}(A_3 \mid A_1 \cap A_2) = \frac{4}{57}$, and so on, down to $\mathbb{P}(A_6 \mid A_1 \cap \cdots \cap A_5) = \frac{1}{54}$.

So, using the chain rule, we get

$$\begin{aligned}&\mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_6)\\ &\qquad = \mathbb{P}(A_1)\,\mathbb{P}(A_2 \mid A_1)\,\mathbb{P}(A_3 \mid A_1 \cap A_2) \cdots \mathbb{P}(A_6 \mid A_1 \cap \cdots \cap A_5)\\ &\qquad = \frac{6}{59} \times \frac{5}{58} \times \frac{4}{57} \times \frac{3}{56} \times \frac{2}{55} \times \frac{1}{54}.\end{aligned}$$

The answer we got before was

$$\frac{6 \times 5 \times 4 \times 3 \times 2 \times 1}{59 \times 58 \times 57 \times 56 \times 55 \times 54}.$$

It's easy to see that this is the same answer, and the structure of the answers shows how the old method got the answer "all at once", while this new method gets the answer "one stage at a time".

4.4 Law of total probability

Example 4.4. *My friend has three dice: a 4-sided dice, a 6-side dice, and a 10-side dice. He picks one of them at random, with each dice equally likely. What is the probability my friend rolls a 5?*

If my friend were to tell which dice he picked, then this question would be very easy! If we write D_4, D_6 and D_{10} to be the events that he picks the 4-sided, 6-sided, or 10-sided dice, then we know immediately that

$$\mathbb{P}(\text{roll } 4 \mid D_4) = 0 \qquad \mathbb{P}(\text{roll } 4 \mid D_6) = \tfrac{1}{6} \qquad \mathbb{P}(\text{roll } 4 \mid D_{10}) = \tfrac{1}{10}.$$

What we need is a way to combine the results for different "sub-cases" into an over-all answer.

Luckily, there exists just such a tool for this job! It's called the "law of total probability" (also known as the "partition theorem"). The important point is to make sure that the different sub-cases cover all possibilities, but that only one of them happens at a time.

Definition 4.3. A set of events $A_1, A_2, \dots, A_n$ are said to be a **partition** of the sample space Ω if

1. they are disjoint, in that $A_i \cap A_j = \emptyset$ for all $i \neq j$;
2. they cover space, in that $A_1 \cup A_2 \cup \cdots \cup A_n = \Omega$.

Theorem 4.2 (Law of total probability)**.** *Let $A_1, A_2, \dots, A_n$ be a partition, and B another event. Then*

$$\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(A_i)\,\mathbb{P}(B \mid A_i).$$

So the law of total probability tells us we can add up the probabilities $\mathbb{P}(B \mid A_i)$ for each of the sub-cases provided we weight them by how likely $\mathbb{P}(A_i)$ by how likely each sub-case is.

Proof. Since the partition of A_is cover space, we can split up B depending on which part of the partition it is in:

$$B = (B \cap A_1) \cup (B \cap A_2) \cup \cdots \cup (B \cap A_n).$$

[I meant to draw a picture here, but didn't get round to it – perhaps you'd like to draw your own?]

Since the A_i are disjoint, the union on the right is disjoint also. Therefore we can use Axiom 3 to get

$$\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(B \cap A_i).$$

But using the definition of conditional probability, each "summand" (term inside the sum) is

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

The result follows. □

Returning to our dice example, D_4, D_6, D_{10} is indeed a partition, since these are the only possibilities and we only choose one dice. So the law of total probability tells us that

$$\mathbb{P}(\text{roll } 5) = \mathbb{P}(D_4)\,\mathbb{P}(\text{roll } 5 \mid D_4) + \mathbb{P}(D_6)\,\mathbb{P}(\text{roll } 5 \mid D_6) + \mathbb{P}(D_{10})\,\mathbb{P}(\text{roll } 5 \mid D_{10}).$$

We were told that all the dice were picked with equal probability, so $\mathbb{P}(D_4) = \mathbb{P}(D_6) = \mathbb{P}(D_{10}) = \frac{1}{3}$, and we calculated the individual conditional probabilities as

$$\mathbb{P}(\text{roll } 4 \mid D_4) = 0 \qquad \mathbb{P}(\text{roll } 4 \mid D_6) = \tfrac{1}{6} \qquad \mathbb{P}(\text{roll } 4 \mid D_{10}) = \tfrac{1}{10}.$$

Therefore, we have

$$\mathbb{P}(\text{roll } 5) = \tfrac{1}{3} \times 0 + \tfrac{1}{3} \times \tfrac{1}{6} + \tfrac{1}{3} \times \tfrac{1}{10} = \tfrac{8}{90} = 0.089.$$

4.5 Bayes' theorem

In this subsection, we will discuss an important result called **Bayes' theorem**. Let's first state and prove this result, and do an example, and then afterwards we'll talk about two reasons why Bayes' theorem is so important.

Theorem 4.3 (Bayes' theorem)**.** *For events A and B with $\mathbb{P}(A), \mathbb{P}(B) > 0$, we have*

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

Bayes' theorem is thought to have first appeared in the writings of Rev. Thomas Bayes, a British church minister and mathematician, shortly after his death, in the 1760s. However, his work was significantly edited by Richard Price, another minister–mathematician, and many people think that Price deserves a large share of the credit.

Proof. From the definition of conditional probability, we can write $\mathbb{P}(A \cap B)$ in two different ways: we can write it as

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

but we can also write it as

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

Since these are two different ways of writing the same thing, we can equate them, to get

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

Dividing both sides by $\mathbb{P}(B)$ gives the result. □

Example 4.5. *My friend again secretly picks the 4-sided, 6-sided, or 10-sided dice, each with probability $\frac{1}{3}$. He rolls that secret dice, and tells me he rolled a 5. What is the probability he picked the 6-sided dice?*

This is asking us to calculate $\mathbb{P}(D_6 \mid \text{roll } 5)$. Bayes' theorem tells us that

$$\mathbb{P}(D_6 \mid \text{roll } 5) = \frac{\mathbb{P}(D_6)\,\mathbb{P}(\text{roll } 5 \mid D_6)}{\mathbb{P}(\text{roll } 5)} = \frac{\frac{1}{3} \times \frac{1}{6}}{\frac{8}{90}} = \frac{5}{8},$$

since we had calculated $\mathbb{P}(\text{roll } 5) = \frac{8}{90}$ in the previous subsection.

The first way to think about Bayes' theorem is that it tells us how to relate $\mathbb{P}(A \mid B)$ and $\mathbb{P}(B \mid A)$. Remember that $\mathbb{P}(A \mid B)$ and $\mathbb{P}(B \mid A)$ are not the same thing! The conditional probability someone is under 40 given they are a Premiership footballer is very high, but the conditional probability someone is a Premiership footballer given they are under 40 is very low.

Bayes' theorem, in this first view, is a useful technical result that helps us switch the order of a conditional probability from B given A to A given B: we have

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

In the dice example, the probability $\mathbb{P}(\text{roll } 5 \mid D_6) = \frac{1}{6}$ was very obvious, but Bayes' theorem allowed us to reverse the conditioning, to find $\mathbb{P}(D_6 \mid \text{roll } 5) = \frac{5}{8}$ instead.

The second way to think about Bayes' rule is that it tells us how to update our beliefs as we acquire more evidence. That is, we might start by believing that the probability some event A will occur is $\mathbb{P}(A)$. But then we find out that B has occurred, so we want to incorporate that knowledge and update our belief of the probability A will occur to $\mathbb{P}(A \mid B)$, the conditional probability A will occur given this new evidence B.

Bayes theorem, in this second view, tells us how to update from $\mathbb{P}(A)$ to $\mathbb{P}(A \mid B)$: we have

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

In the dice example, we initially believed there was a $\mathbb{P}(D_6) = \frac{1}{3} = 0.333$ chance our friend had chosen the six-sided dice. But when we heard that our friend had rolled a 5, we updated our belief to now thinking there was now a $\mathbb{P}(D_6 \mid \text{roll } 5) = \frac{5}{8} = 0.625$ chance it was the 6-sided dice.

This second way of thinking about Bayes' theorem is at the heart of **Bayesian statistics**. In Bayesian statistics, we start with a "prior" belief about a model, then, after collecting some data, we update to a "posterior" belief, according to the rules of Bayes' theorem. We will discuss Bayesian statistics much more in Section 10.

Quite often we use Bayes' theorem and the law of total probability together. If we have a partition $A_1, A_2, \dots, A_n$, perhaps representing some possible hypotheses, and we observe an event B, then Bayes' theorem tells us how likely each

hypothesis is given the observation:

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

But this shared denominator $\mathbb{P}(B)$ can be expanded using the law of total probability

$$\mathbb{P}(B) = \sum_{j=1}^{n} \mathbb{P}(A_j)\,\mathbb{P}(B \mid A_j).$$

Together, we get the following.

Theorem 4.4. *Let $\{A_1, A_2, \dots, A_n\}$ be a partition of a sample space and let B be another event. Then, for all $i = 1, 2, \dots, n$, we have*

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

This is essentially what we did with the dice example, although we split it up into two separate parts rather than using this formula directly.

4.6 Diagnostic testing

Members of the public are tested for a certain disease. About 2% of the population have the disease. The test is 95% accurate, in the following sense: if you have the disease, there's a 95% chance you correctly get a positive test result, while if you don't have the disease, there's a 95% chance you correctly get a negative test result. Suppose you get a positive test result. What is the probability you have the disease?

The first thing we have to do is translate the words in the question into probability statements. Let D be the event you have the disease, so D^{c} is the event you don't have the disease, and let $+$ be the event you get a positive result. Then the question tells us that

- $\mathbb{P}(D) = 0.02$ and $\mathbb{P}(D^{\mathsf{c}}) = 0.98$;
- $\mathbb{P}(+ \mid D) = 0.95$ and $\mathbb{P}(+ \mid D^{\mathsf{c}}) = 0.05$;
- we want to find $\mathbb{P}(D \mid +)$.

Note also that D (you have the disease) and D^{c} (you don't) make up a partition. Then Theorem 4.4 tells us that

$$\mathbb{P}(D \mid +) = \frac{\mathbb{P}(D)\,\mathbb{P}(+ \mid D)}{\mathbb{P}(D)\,\mathbb{P}(+ \mid D) + \mathbb{P}(D^{\mathsf{c}})\,\mathbb{P}(+ \mid D^{\mathsf{c}})}.$$

Putting in all the numbers we have, we get

$$\mathbb{P}(D \mid +) = \frac{0.02 \times 0.95}{0.02 \times 0.95 + 0.98 \times 0.05} = 0.28.$$

So if you get a positive result on this 95%-accurate test, there's still only about a 1 in 4 chance you actually have the disease.

Many people find this result surprising. It sometimes helps to put more concrete numbers on things. Suppose 1000 people get tested. On average, we expect about 20 of them to have the disease, and 980 of to not have the disease. Of the 20 with the disease, on average 19 will correctly test positive, while 1 will test negative. Of the 980 without the disease, an average 931 will correctly test negative, but 49 will wrongly test positive. So of the $19 + 49 = 68$ people with positive tests, only 19 of them actually have the disease, which is 28%.

The key point is that the disease is rare – only 2% of people have it. So even though positive test increases the likelihood you have the disease a lot (it's about 14 times more likely), it's not enough to make it a very large probability.

Summary

- Two events are independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$.
- The conditional probability of B given A is $\mathbb{P}(B \mid A) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}$.
- The law of total probability says that if $A_1, A_2, \ldots A_n$ is a partition of the sample space, then

$$\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(A_i)\,\mathbb{P}(B \mid A_i).$$

- Bayes' theorem says that $\mathbb{P}(A \mid B) = \dfrac{\mathbb{P}(A)\,\mathbb{P}(B \mid A)}{\mathbb{P}(B)}$.

Chapter 5

Discrete random variables

5.1 What is a random variable?

Let's consider again the case of rolling two dice. We know that the sample space is the set of pairs of numbers between 1 and 6

$$\Omega = \{\omega = (\omega_1, \omega_2) : \omega_1, \omega_2 \in \{1, 2, 3, 4, 5, 6\}\},$$

which is equipped with the classical probability measure

$$\mathbb{P}(A) = \frac{|A|}{36}.$$

But if we are rolling the two dice as part of a board game, we might only care about the total score on the two dice, rather than the two individual dice scores. In this case, we could write X for the score on the first dice plus the score on the second dice – that is,

$$X(\omega) = \omega_1 + \omega_2.$$

If we want to know the probability we roll a total of 7, say, then we could write this as $\mathbb{P}(A)$ where

$$A = \{\omega : \omega_1 + \omega_2 = 7\}$$

is the set of dice rolls with total 7. But it might just be easier to write this as $\mathbb{P}(X = 7)$. We might also be interested in other things about the total score X, like what the average total score is.

Here, X is an example of what we call a **random variable**. A random variable can be thought of as a numerical summary of an experiment (like the total summarising the two individual dice rolls). But by considering the random variable directly, it often means we don't have to worry so much about exactly what the sample space is, or what probability measure is being used, or which sample outcomes are in a particular event. This often makes our life easier when thinking about probability problems.

Random variables are typically given capital letters from late in the alphabet, like X, Y, Z. Values that those random variables take are often given lower-case letters, like x, y, z.

The formal definition of a random variable is as a function that turns the outcome into this numerical summary.

Definition 5.1. Let Ω be a sample space. Then a **random variable** is a function X from Ω to the real numbers $\mathbb{R}$; that is, to each sample outcome ω it assigns a real number $X(\omega)$.

Expressions like $\mathbb{P}(X = x)$ should be understood as representing more formal probability

$$\mathbb{P}\big(\{\omega : X(\omega) = x\}\big).$$

This formal definition of a random variable as a function was summarised by my own first-year probability lecturer as "There's only two things you need to know about the definition of a random variable: first, it's not random; second, it's not a variable."

However, the way we actually *think* about random variables *is* as random and is as variables. In this more informal way of thinking, a random variable is a variable that can take different values with different probabilities – just as the total of the two dice can be 2 with probability $\frac{1}{36}$, or the value 3 with probability $\frac{2}{36}$, and so on.

Definition 5.2. The set of values a random variable X can take is called its **range**, $\text{Range}(X) = \{X(\omega) : \omega \in \Omega\}$.

So, for example, the range of the dice sum X is $\text{Range}(X) = \{2, 3, \dots, 12\}$.

Random variables that we will consider in this module will be one of two types:

- **Discrete random variables** have a range that is finite (like the dice total being an integer between 2 and 12) or countably infinite (like the positive integers, for example). Discrete random variables can be used as models for "count data".
- **Continuous random variables** have a range that is uncountably infinite (like the real numbers, the positive real numbers, or the interval $[0, 1]$, for example). Continuous random variables can be used as models for "measurement data".

In this section and the next two, we will look at discrete random variables; later in Section 8 and 9 we will look at continuous random variables.

5.2 Probability mass functions

We now consider only discrete random variables X, where the range $\text{Range}(X)$ is a finite or countably infinite set. In this case, Axiom 3 tells us that for any set A, we have

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

(Recall that the symbol $\in$ means "is an element of", or just "is in" for short.) So to fully understand a discrete random variable X, we need only understand the probabilities $\mathbb{P}(X = x)$. These are captured by the probability mass function.

Definition 5.3. For a discrete random variable X, its **probability mass function** (or **PMF**) is the function p_X where

$$p_X(x) = \mathbb{P}(X = x) \qquad \text{for } x \in \text{Range}(X).$$

(When the random variable is obvious from context, we'll just write $p(x)$ without the subscript.)

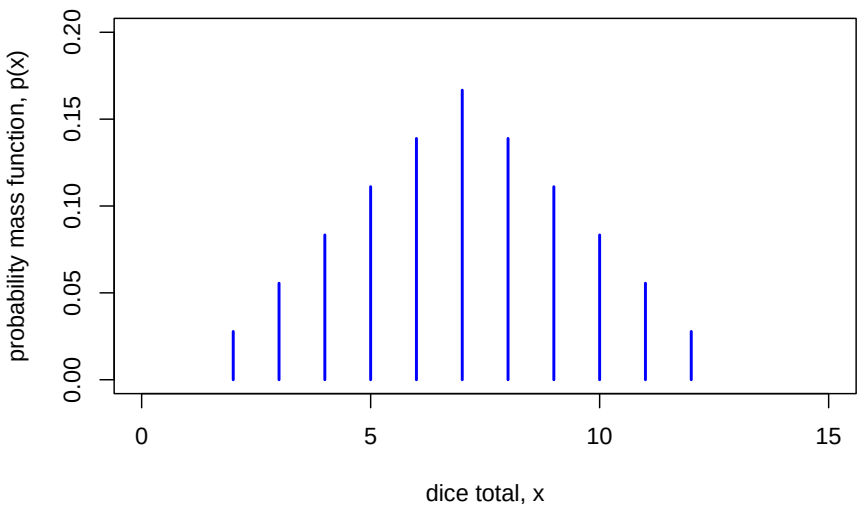
The key is that, once we are thinking in terms of a random variable via its PMF, we can (usually) stop worrying to much about what the underlying sample space is and how the random variable acts on that space.

Example 5.1. Let X being the sum of two dice rolls. As this is a classical probability problem, the probability $p(x)$ of rolling a total of x is $n(x)/36$, where $n(x)$ is the number of ways of rolling a total of x. So, for example, there is only one way $(1, 1)$ of rolling a total of 2, so $p(2) = \frac{1}{36}$, but there are 5 ways of rolling a 6: $(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)$; so $p(5) = \frac{5}{36}$.

The PMF p of X is given by

x	2	3	4	5	6	7
$p(x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$

x	$\cdots$	8	9	10	11	12
$p(x)$	$\cdots$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$



Example 5.2. Consider tossing a biased coin, that is Heads with probability p and Tails with probability $1-p$. Let $X = 1$ if the coin lands Heads, and $X = 0$ if the coin lands Tails. The PMF p_X of this random variable is given by

$$p_X(0) = 1 - p \qquad p_X(1) = p.$$

We could alternatively think of the same random variable as representing the result of an experiment, where $X = 1$ represents a success, with probability $p_X(1) = p$, and $X = 0$ represents a failure, with probability $p_X(0) = 1 - p$.

A random variable X with this PMF is called a **Bernoulli trial** (or a "Bernoulli random variable", or is said to "follow the Bernoulli distribution" – after the seventeenth-century Swiss mathematician Jacob Bernoulli). We use the notation $X \sim \mathrm{Bern}(p)$ for short.

Note that since $p(x) = \mathbb{P}(X = x)$ is a probability, we must have $p(x) \geq 0$ for all $x \in \mathrm{Range}(X)$. Further, we have

$$1 = \mathbb{P}(\Omega) = \mathbb{P}\big(X \in \mathrm{Range}(X)\big) = \sum_{x \in \mathrm{Range}(X)} \mathbb{P}(X = x) = \sum_{x \in \mathrm{Range}(X)} p(x).$$

Hence we have the following:

Theorem 5.1. *Let X be a discrete random variable, and let p_X be its PMF. Then*

- $p_X(x) \geq 0$ *for all* x*;*
- $\displaystyle\sum_x p_X(x) = 1.$

Sometimes it is useful to know the probability a random variable X is less than some value x. This is captured by the **cumulative distribution function** (or **CDF**) F_X, where

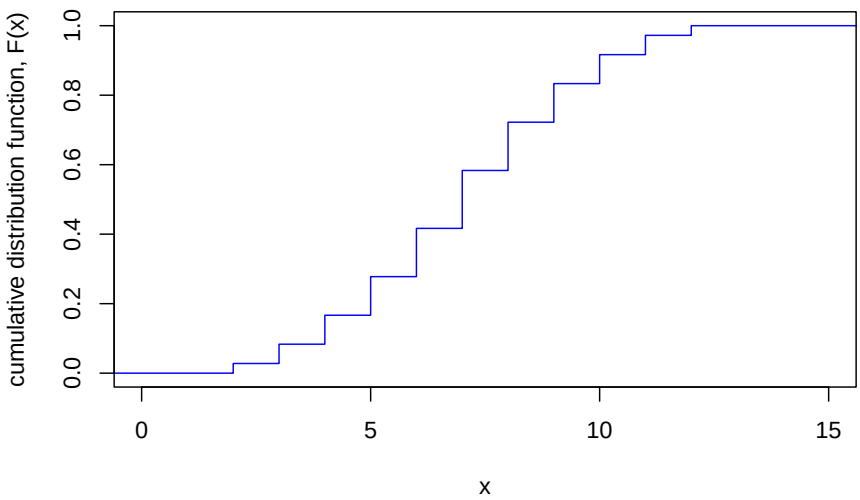
$$F_X(x) = \mathbb{P}(X \leq x) = \sum_{y \leq x} p_X(y) \qquad \text{for } x \in \mathbb{R}.$$

Example 5.3. Let $X \sim \mathrm{Bern}(p)$ be a Bernoulli random variable with success probability p. Then its CDF F is

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 - p & \text{for } 0 \leq x < 1 \\ 1 & \text{for } x \geq 1. \end{cases}$$

Example 5.4. If X is the sum of two dice rolls, then the CDF F is given by

$x \in$	$(-\infty, 2)$	$[2, 3)$	$[3, 4)$	$[4, 5)$	$\cdots$	$[11, 12)$	$[12, \infty)$
$F(x)$	0	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$\cdots$	$\frac{35}{36}$	1



Note that the CDF is a "step function" that starts at 0, then jumps up suddenly at each of the values $2, 3, \ldots, 12$, ending up at 1.

For any random variable X with CDF F,

- if x is smaller than everything in the range of X, then $F(x) = 0$;
- if x is greater than everything in the range of X, then $F(x) = 1$;
- $F(x)$ is increasing in x.

5.3 Expectation

Often, we will be interested in the "average" value of a random variable – for example, the average total from two dice rolls – which represents what "central" values of the random variable. This average is called the "expectation".

Definition 5.4. Let Ω be a finite or countably infinite sample space, $\mathbb{P}$ be a probability measure on Ω, and X be a discrete random variable on Ω. Then the **expectation** (or **expected value**) of X is

$$\mathbb{E}X = \sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\{\omega\}).$$

If p is the PMF of X, then a more convenient formula is

$$\mathbb{E}X = \sum_{x \in \mathrm{Range}(X)} x\, p_X(x).$$

We get the second formula from the first by grouping together all outcomes ω that lead to the same value $x = X(\omega)$ of X.

Note that "expectation" is simply the name that mathematicians give to the value $\mathbb{E}X = \sum_x x\,p(x)$. We don't necessarily "expect" to get the value $\mathbb{E}X$ as the outcome in the normal English-language sense of the word "expect". (Indeed, you might like to check that the expectation of a single dice roll is 3.5, but you certainly don't "expect" to get the number 3.5 in a single roll of the dice!) We will see later that the the expectation can be interpreted as a sort of "long-run mean outcome".

Example 5.5. *Let $X \sim Bern(p)$ be a Bernoulli trial with success probability p. What is the expectation $\mathbb{E}X$?*

Using the second formula in the definition, we have

$$\mathbb{E}X = \sum_x x\,p(x) = 0 \times (1-p) + 1 \times p = p.$$

Example 5.6. When X is the total of two dice rolls, the expectation is

$$\begin{aligned}\mathbb{E}X &= \sum_{x \in \mathrm{Range}(X)} x\,p(x) \\ &= 2 \times \tfrac{1}{36} + 3 \times \tfrac{2}{36} + \cdots + 12 \times \tfrac{1}{36} \\ &= \tfrac{252}{36} \\ &= 7.\end{aligned}$$

5.4 Functions of random variables

In previous examples, we looked at X being the total of the dice rolls. But we could equally well chosen to have looked at a different random variable that is a function of that total X, like "double the total and add 1" $Y = 2X + 1$, or "the total minus 4, all squared" $Z = (X-4)^2$. (I'm not sure *why* you'd care about these, but you could study them if you wanted to…)

Example 5.7. Let $Y = 2X + 1$. Then for each potential outcome x of X, there is a matching outcome $y = 2x + 1$ of Y. So we can find the PMF for Y by keeping the same probabilities as for X, but changing the values x to the values $y = 2x + 1$.

y	5	7	9	$\cdots$	23	25
$p_Y(y)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\cdots$	$\frac{2}{36}$	$\frac{1}{36}$

Example 5.8. What about $Z = (X-4)^2$? This is a bit trickier, because more than one outcome x can lead to the same outcome $z = (x-4)^2$: for example, $(2-4)^2 = (6-4)^2 = 4$. So, more precisely, we have

$$\mathbb{P}(Z = z) = \mathbb{P}(X \in \{x : (x-4)^2 = z\}),$$

or, in terms of PMFs,

$$p_Z(z) = \sum_{x:(x-4)^2=z} p_X(x).$$

z	0	1	4	$\cdots$	49	64
$p_Z(z)$	$\frac{3}{36}$	$\frac{2}{36} + \frac{4}{36} = \frac{6}{36}$	$\frac{1}{36} + \frac{5}{36} = \frac{6}{36}$	$\cdots$	$\frac{2}{36}$	$\frac{1}{36}$

So if we wanted to find the expectation of a function of a random variable, we could first find the PMF, like in the above examples, and then use that PMF to find the expectation. But there is a quicker way.

Theorem 5.2 (Law of the unconscious statistician)**.** *Let X be a random variable, and let $Y = g(X)$ be another random variable that is a function of X. Then*

$$\mathbb{E}Y = \mathbb{E}g(X) = \sum_x g(x)\, p_X(x).$$

(The rather cruel name of this theorem is, I think, because this is the formula you might carelessly write down for $\mathbb{E}g(X)$ if you weren't thinking carefully – but it turns out it's correct!)

Proof. As in the previous example, the idea is to group together xs that give the same y.

From the definition of expectation, we have

$$\mathbb{E}Y = \sum_y y\, p_Y(y).$$

Then using

$$p_Y(y) = \sum_{x:g(x)=y} p_X(x),$$

we get

$$\begin{aligned}\mathbb{E}Y &= \sum_y y \sum_{x:g(x)=y} p_X(x) \\ &= \sum_y \sum_{x:g(x)=y} y\, p_X(x) \\ &= \sum_y \sum_{x:g(x)=y} g(x)\, p_X(x),\end{aligned}$$

since $y = g(x)$ inside the second sum. But those two sums together are summing over all x, just partitioned by which value of y they lead to, so can be replaced by a single sum over x. That gives the theorem. □

There are some functions for which this expression becomes particularly simple.

Theorem 5.3 (Linearity of expectation, 1)**.** *Let X be a random variable. Then*

1. $\mathbb{E}(aX) = a\mathbb{E}X$;
2. $\mathbb{E}(X + b) = \mathbb{E}X + b$.

Proof. We use the law of the unconscious statistician.

For part 1, we can take the a outside the sum, to get

$$\mathbb{E}(aX) = \sum_x ax\, p_X(x) = a \sum_x x\, p_X(x) = a\mathbb{E}X.$$

For part 2, we have

$$\begin{aligned}
\mathbb{E}(X+b) &= \sum_x (x+b)\, p_X(x) \\
&= \sum_x \left(x\, p_X(x) + b\, p_X(x)\right) \\
&= \sum_x x\, p_X(x) + \sum_x b\, p_X(x) \\
&= \mathbb{E}(X) + b \sum_x p_X(x) \\
&= \mathbb{E}(X) + b.
\end{aligned}$$

The last line was because PMFs always add up to 1, so $\sum_x p_X(x) = 1$. □

So for our "double the dice total and add 1" random variable $Y = 2X + 1$, we have

$$\mathbb{E}Y = \mathbb{E}(2X+1) = 2\mathbb{E}X + 1 = 2 \times 7 + 1 = 15.$$

5.5 Variance

In the same way as the expectation of a random variable tells us about central typical values of it, the variance of a random variable tells us about the spread of typical values.

Definition 5.5. Let X be a random variable with expectation $\mathbb{E}X = \mu$. Then the **variance** of X is

$$\mathrm{Var}(X) = \mathbb{E}(X-\mu)^2.$$

(To be clear, the notation there means the expectation of $(X-\mu)^2$; and *not* $\mathbb{E}(X-\mu)$ squared, which would be $0^2 = 0$.)

Note that $(X-\mu)^2$ is a square, so always non-negative, and hence the variance is always non-negative also.

It may not surprise you, if you remember Section 1 that to go along with that definitional formula for the variance, we also have a computational formula.

Theorem 5.4. *Let X be a random variable with expectation $\mathbb{E}X = \mu$. Then the variance $\mathrm{Var}(X) = \mathbb{E}(X-\mu)^2$ can also be calculated as*

$$\mathrm{Var}(X) = \mathbb{E}X^2 - \mu^2.$$

(Again, $\mathbb{E}X^2$ means the expectation of X^2.)

Proof. As previously we expand out the brackets, and use linearity of expectation (in the same way we "brought the sum inside" previously). We get

$$\begin{aligned}\operatorname{Var}(X) &= \mathbb{E}(X-\mu)^2 \\ &= \mathbb{E}(X^2 - 2\mu X + \mu^2) \\ &= \mathbb{E}X^2 - \mathbb{E}(2\mu X) + \mathbb{E}\mu^2 \\ &= \mathbb{E}X^2 - 2\mu\,\mathbb{E}X + \mu^2.\end{aligned}$$

But we said that $\mathbb{E}X$ would be called μ, so we can substitute in $\mathbb{E}X = \mu$, to get

$$\operatorname{Var}(X) = \mathbb{E}X^2 - 2\mu^2 + \mu^2 = \mathbb{E}X^2 - \mu^2,$$

as required. □

Example 5.9. Let $X \sim \operatorname{Bern}(p)$ be a Bernoulli trial, and recall that $\mathbb{E}X = p$. Using the definitional formula, we have

$$\begin{aligned}\operatorname{Var}(X) &= \mathbb{E}(X-p)^2 \\ &= (0-p)^2\, p_X(0) + (1-p)^2\, p_X(1) \\ &= p^2 \times (1-p) + (1-p)^2 \times p \\ &= p(1-p)(p + (1-p)) \\ &= p(1-p).\end{aligned}$$

Alternatively, using the computational formula, we have

$$\begin{aligned}\operatorname{Var}(X) &= \mathbb{E}X^2 - p^2 \\ &= (0^2\, p_X(0) + 1^2 p_X(1)) - p^2 \\ &= 0 \times (1-p) + 1 \times p - p^2 \\ &= p - p^2 \\ &= p(1-p).\end{aligned}$$

Example 5.10. For the total of two dice, using the computational formula, we have

$$\begin{aligned}\operatorname{Var}(X) &= \mathbb{E}X^2 - \mu^2 \\ &= \left(2^2 \times \frac{1}{36} + 3^2 \times \frac{2}{36} + \cdots + 12^2 \times \frac{1}{36}\right) - 7^2 \\ &= \frac{1974}{36} - 49 \\ &= \frac{70}{12} \approx 5.8.\end{aligned}$$

Finally, a result on what happens to the variance of simple functions of random variables.

Theorem 5.5. *Let X be a random variable. Then*

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

You will prove this on the problem sheet.

Summary

- A random variable is a numerical summary of a random experiment.
- The probability mass function (PMF) is $p_X(x) = \mathbb{P}(X = x)$, and the cumulative distribution function (CDF) is $F_X(x) = \mathbb{P}(X \leq x)$.
- The expectation is $\mathbb{E}X = \sum_x x\, p_X(x)$.
- The variance of a random variable with expectation μ is $\mathrm{Var}(X) = \mathbb{E}(X - \mu)^2$.
- $\mathbb{E}(aX + b) = a\mathbb{E}X + b$ and $\mathrm{Var}(aX + b) = a^2\,\mathrm{Var}(X)$.

Problem Sheet 3

This is Problem Sheet 3. This problem sheet covers Sections 4 and 5. You should work through all the questions on this problem sheet during Weeks 4 and 5, in preparation for your tutorial in Week 6. The problem sheet contains two assessed questions, which are due in by **2pm on Monday 8 November**.

A: Short questions

A1. Consider dealing two cards (without replacement) from a pack of cards. Which of the following pairs of events are independent?

(a) "The first card is a Heart" and "The first card is Red".

(b) "The first card is a Heart" and "The first card is a Spade".

(c) "The first card is a Heart" and "The first card is an Ace".

(d) "The first card is a Heart" and "The second card is a Heart".

(e) "The first card is a Heart" and "The second card is an Ace".

A2. Three events A, B, C are said to be *pairwise independent* if each pair of events are independent – that is, if A and B are independent, B and C are independent, and A and C are independent. The three events are said to *mutually independent* if they are pairwise independent and also $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\,\mathbb{P}(B)\,\mathbb{P}(C)$.

Consider rolling two dice independently. Let A be the event that the first roll is even, let B be the event that the second roll is even, and let C be the event that the total score is even.

(a) Are A, B, and C pairwise independent?

(b) Are A, B, and C mutually independent?

A3. Consider the random variable X with the following PMF:

x	-1	0	0.5	1	2
$p(x)$	0.1	0.3	0.3	0.2	0.1

Find the expectation and variance of X.

A4. Consider the random variable X with the following PMF:

x	1	2	4	5	a
$p(x)$	0.1	0.2	0.1	b	c

This random variable has $\mathbb{E}X = 4.3$ and $\mathrm{Var}(X) = 4.61$. Find the values of a, b and c.

A5. The temperature T_C measured in degrees Celsius can be converted to a temperature T_F in degrees Fahrenheit using the formula $T_F = \frac{9}{5}T_C + 32$.

The average daily maximum temperature in Leeds in July is 19.0 °C with a variance of 10.4 °C^2.

(a) What is the average daily maximum temperature in degrees Fahrenheit?

(b) What is the variance of the daily maximum temperature in degrees Fahrenheit squared?

(c) The average temperature in March is 1.8 °C higher than in February. What is this average difference in degrees Fahrenheit?

B: Long questions

B1. Suppose A and B are a pair of independent events. Show that A and B^c are also independent events.

B2. Chloe cycles to work with probability 0.6 and takes the bus with probability 0.4. She has noticed that she is late 20% of the time when she takes the bus. Her boss notices that, on average, Chloe is late one time per week (that is, once every five days).

(a) What is the probability that Chloe will be late if she cycles to work?

(b) Chloe is late today, and her boss suspects this is because she took the bus. What is the probability that this guess is correct?

*[**Note:** Because $0.2 = \frac{1}{5}$, both types of transport are equally (un)reliable, making this a kind of boring question, which I didn't notice when copying the question from last year's problem sheet. This is a note to remind myself to change the numbers next year.]*

B3. Let Ω be a sample space, let $\mathbb{P}$ be a probability measure on Ω, and fix an event $B \subset \Omega$ with $\mathbb{P}(B) > 0$. Show that the conditional probability $\mathbb{P}(\cdot \mid B)$ is also a probability measure on Ω. That is, show that the conditional probability also satisfies the probability axioms:

1. $\mathbb{P}(A \mid B) \geq 0$ for all events $A \subset \Omega$;
2. $\mathbb{P}(\Omega \mid B) = 1$;
3. For disjoint events $A_1, A_2, \ldots$, we have

$$\mathbb{P}(A_1 \cup A_2 \cup \cdots \mid B) = \mathbb{P}(A_1 \mid B) + \mathbb{P}(A_2 \mid B) + \cdots.$$

B4. Soldiers are asked about their use of illegal drugs, but to protect their privacy and ensure honest answers, a "randomised survey" is used. Each soldier is handed a deck of three cards, picks one of the three cards at random, and responds according to what the card says. The three cards say:

1. "Say 'Yes.' "
2. "Say 'No.' "
3. "Truthfully answer the question 'Have you taken any illegal drugs in the past 12 months?' "

(a) 40% of soldiers respond "Yes". What is the likely proportion of soldiers who have taken illegal drugs in the past 12 months.

(b) If a soldier responds "Yes", what is the probability that the soldier has taken illegal drugs in the past 12 months.

B5. Let X be a random variable. Prove that

(a) $\mathrm{Var}(aX) = a^2 \mathrm{Var}(X)$;

(b) $\mathrm{Var}(X + b) = \mathrm{Var}(X)$.

B6. A gambling game works as follows. You keep tossing a fair coin until you first get a Head. If your first Head comes on the nth coin toss, then you win 2^n pounds.

(a) What is the probability that the first Head is seen on the nth toss of the coin?

(b) Show that the expected winnings from playing this game are infinite.

(c) The "St Petersburg paradox" refers to the fact that, despite the fact that the expected winnings from this game are infinite, hardly anybody would be prepared to play this game for, say, £100, and certainly not for £1000. Discuss a few possible "resolutions" to this paradox which could explain why people are unwilling to play this game despite seemingly having infinite expected winnings.

C: Assessed questions

The last two questions are **assessed questions**. These two questions count for 3% of your final mark for this module.

The deadline for submitting your solutions is **2pm on Monday 8 November** at the beginning of Week 7, although I strongly recommend completing and submitting your work during Week 6. Submission will be via Gradescope; submission will open on Monday 1 November. Your work will be marked by your tutor and returned on Monday 15 November, when solutions will also be made available.

Both questions are "long questions", where the marks are not only for mathematical accuracy but also for the clarity and completeness of your explanations.

You should not collaborate with others on the assessed questions: your answers must represent solely your own work. The University's rules on academic integrity – and the related punishments for violating them – apply to your work on the assessed questions.

C1. A computer spam filter is 98% effective at sending spam emails to my junk folder, but will also incorrectly send 1% of legitimate emails to my junk folder. Suppose that 1 in 10 emails are spam. What proportion of emails in my junk folder are actually legitimate emails?

C2. A random variable X is said to follow the *discrete uniform distribution* on $\{1, 2, \dots, n\}$ if each value x in that set $\{1, 2, \dots, n\}$ is equally likely.

(a) Show that the expectation of X is $\mathbb{E}X = \dfrac{n+1}{2}$.

(b) Find the variance of X.

(c) Let Y be a discrete uniform distribution on $\{a, a+1, a+2, \dots, b-1, b\}$, for integers a and b with $a < b$. Using parts (a) and (b), but without calculating any sums directly, find the expectation and variance of Y.

You may use without proof the standard results

$$\sum_{x=1}^{n} x = \frac{n(n+1)}{2} \qquad \sum_{x=1}^{n} x^2 = \frac{n(n+1)(2n+1)}{6}.$$

Solutions to short questions

A1. (c) and (e) are independent. **A2.** (a) Yes (b) No **A3.** 0.45 and 0.5725 *[corrected]* **A4.** $a = 9, b = 0.5, c = 0.1$ **A5** (a) 66.2 °F (b) 33.7 °F^2 (c) 3.2 °F

Chapter 6

Discrete distributions

In the last section, we developed the idea of random variables, and in particular discrete random variables. We saw that the benefit of random variables is that we can just worry about their distribution, which often allows us to move the sample space Ω and other more technical matters into the background. (Here, we informally use the word "distribution" to refer to the probability mass function of a random variable – or, later, the continuous equivalent, the probability density function).

There are some distributions – or, rather, some families of distributions – that are so useful that we often want to use them for modelling real-world quantities. In this section, we will look at a number of useful discrete distributions.

6.1 Binomial distribution

One family of distributions we have already seen is the Bernoulli trial $\text{Bern}(p)$, which is 1 with probability p and 0 with probability $1-p$. We saw that this could model whether or a biased coin lands Heads, or more generally whether an experiment is successful.

Example 6.1. *Suppose we toss 10 independent biased coins, each of which lands Heads with probability 0.7 and Tails with probability 0.3. What is the probability we get exactly 8 Heads altogether?*

The probability that any specific 8 coins land Heads and the other 2 land Tails is $0.7^8 \times 0.3^2$. However, there are $\binom{10}{8}$ choices for which 8 coins are the ones that land Heads. Hence, the probability is

$$\mathbb{P}(8 \text{ Heads}) = \binom{10}{8} \times 0.7^8 \times 0.3^2 = 0.23.$$

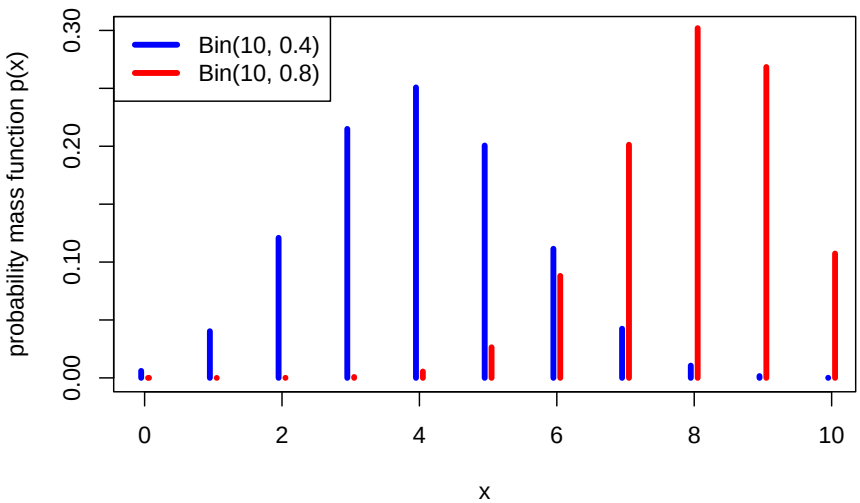
This is a special case of the binomial distribution.

Definition 6.1. Let X be a discrete random variable with range $\{0, 1, 2, \dots, n\}$ and PMF

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}.$$

Then we say that X follows the **binomial distribution** with parameters n and k, and write $X \sim \text{Bin}(n, p)$.

So a binomial random variable represents the number of successes in n Bernoulli trials. In our previous example, the number of Heads from the coin tosses was $\text{Bin}(10, 0.7)$.



Theorem 6.1. *Let $X \sim Bin(n, p)$. Then*

- $\mathbb{E}X = np$,
- $\text{Var}(X) = np(1-p)$.

You can prove this by working out the sums by hand – for example, the expectation is the value of the sum

$$\mathbb{E}X = \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x}.$$

However, in next section we will see an easier way, so we'll reserve the proof until then.

So for my 10 biased coins, that are each Heads with probability 0.7, the expectation and variance are

$$\mathbb{E}X = 10 \times 0.7 = 7$$
$$\text{Var}(X) = 10 \times 0.7 \times 0.3 = 2.1$$

6.2 Geometric distribution

Example 6.2. *I decide to roll a fair dice until I first roll a six, and then stop. What's the probability I get the first six on my 5th roll of the dice?*

For the first six to be on the 5th attempt, the first 4 rolls have to be non-sixes, and then the fifth roll has to be a six. This has probability

$$\left(\tfrac{5}{6}\right)^4 \times \tfrac{1}{6} = \tfrac{625}{7776} = 0.08.$$

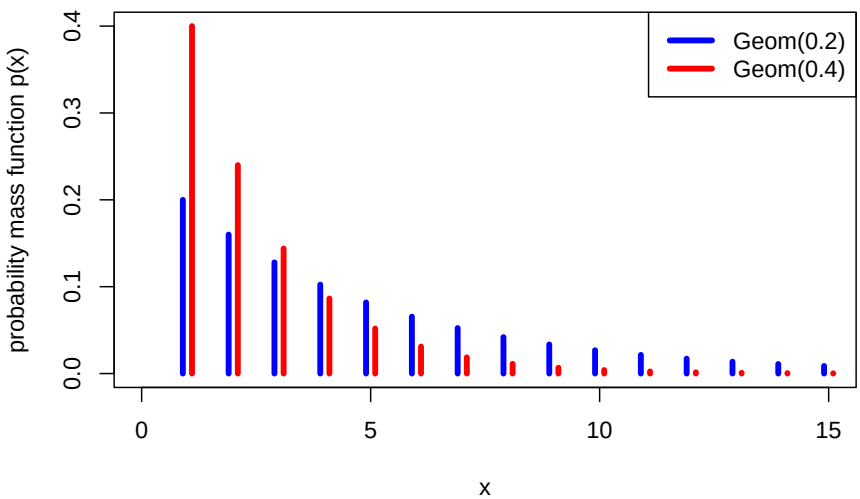
This is a special case of the geometric distribution.

Definition 6.2. Let X be a discrete random variable with range $\{1, 2, \dots\}$ and PMF

$$p(x) = (1-p)^{x-1}p.$$

Then we say that X follows the **geometric distribution** with parameter p, and write $X \sim \text{Geom}(p)$.

So a geometric random variable represents the number Bernoulli trials until the first success. In our previous example, the number of dice rolls until a six was $\text{Geom}(\frac{1}{6})$.



Theorem 6.2. *Let $X \sim Geom(n, p)$. Then*

- $\mathbb{E}X = \dfrac{1}{p}$,
- $\text{Var}(X) = \dfrac{1-p}{p^2}$.

So the expected number of rolls until rolling a six is

$$\mathbb{E}X = \frac{1}{\frac{1}{6}} = 6,$$

with variance

$$\mathrm{Var}(X) = \frac{1 - \frac{1}{6}}{\left(\frac{1}{6}\right)^2} = 30.$$

Proof. (Non-examinable) For the expectation, we want to calculate

$$\mathbb{E}X = \sum_{x=1}^{\infty} x(1-p)^{x-1}p = p\sum_{x=0}^{\infty} x(1-p)^{x-1}.$$

(We can include the $x = 0$ term in the sum since it is equal to 0.)

At this point we will invoke the identity

$$\sum_{x=0}^{\infty} xa^{x-1} = \frac{1}{(1-a)^2},$$

which can be proved by differentiating the standard sum of a geometric progression

$$\sum_{x=0}^{\infty} a^x = \frac{1}{1-a}$$

with respect to a.

Using that identity with $a = 1 - p$, we get

$$\mathbb{E}X = p\sum_{x=0}^{\infty} x(1-p)^{x-1} = p\,\frac{1}{\big(1-(1-p)\big)^2} = \frac{1}{p},$$

as required.

For the variance, we will use a trick that sometimes comes in useful, which is to start by calculating $\mathbb{E}X(X-1)$. Here we get

$$\mathbb{E}X(X-1) = \sum_{x=1}^{\infty} x(x-1)(1-p)^{x-1}p = p(1-p)\sum_{x=0}^{\infty} x(x-1)(1-p)^{x-2}.$$

To calculate the sum, we note that differentiating the geometric progression formula twice gives

$$\sum_{x=0}^{\infty} x(x-1)a^{x-2} = \frac{2}{(1-a)^3},$$

so we get

$$\mathbb{E}X(X-1) = p(1-p)\sum_{x=0}^{\infty} x(x-1)(1-p)^{x-2} = p(1-p)\,\frac{2}{p^3} = \frac{2(1-p)}{p^2}.$$

We now want to use the computational formula $\mathrm{Var}(X) = \mathbb{E}X^2 - \mu^2$ to get the variance. We know $\mu = 1/p$, and from the calculation above, we have

$$\mathbb{E}X(X-1) = \mathbb{E}X^2 - \mathbb{E}X = \mathbb{E}X^2 - \frac{1}{p} = \frac{2(1-p)}{p^2}.$$

So

$$\begin{aligned}\mathrm{Var}(X) = \mathbb{E}X^2 - \mu^2 &= \left(\frac{2(1-p)}{p^2} + \frac{1}{p}\right) - \left(\frac{1}{p}\right)^2 \\ &= \frac{2(1-p)+p-1}{p^2} \\ &= \frac{1-p}{p^2}.\end{aligned}$$

□

Note: Here, we defined a geometric random variable as being the number of trials up to and including the first success, which is a number in $\{1, 2, \dots\}$. However, some authors define it as the number of failures *before* the first success, which is a number in $\{0, 1, 2, \dots\}$. If X is our definition and Y is the second "number of failures" definition, then X and $Y+1$ have the same distribution. Annoyingly, R uses the "number of failures" definition, as we will discuss in a later R worksheet.

6.3 Poisson distribution

Another important distribution is the Poisson distribution. The Poisson distribution (roughly "*pwa*-song") is typically used to model "the number of times something happens in a set period of time". For example, the number of emails you receive in a day; the number of claims at an insurance company each year; or the number of calls to call centre in one hour. (Famously, one of the first historical datasets modelled using a Poisson distribution was "the number of Prussian soldiers in different cavalry units kicked to death by their horse between 1875 and 1894".) We'll explain why the Poisson distribution is a good model for this in the next subsection.

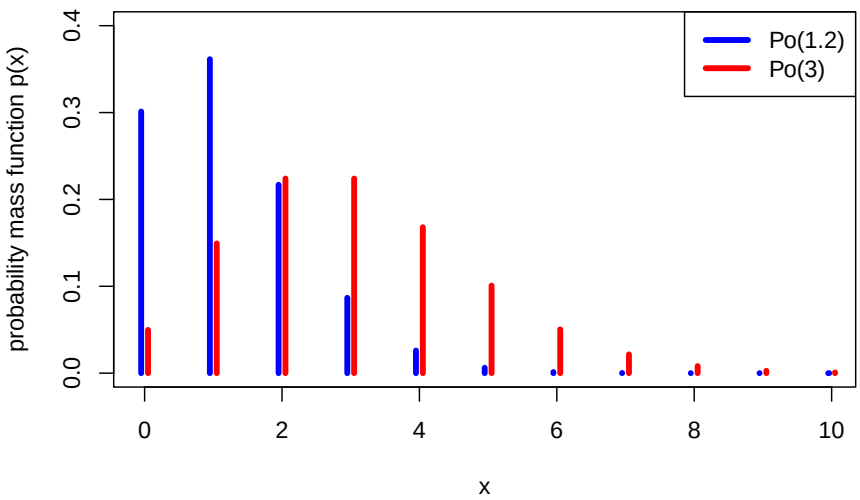
Definition 6.3. Let X be a discrete random variable with range $\{0, 1, 2, \dots\}$ and PMF

$$p(x) = \mathrm{e}^{-\lambda}\frac{\lambda^x}{x!}.$$

Then we say that X follows the **Poisson distribution** with **rate** λ, and write $X \sim \mathrm{Po}(\lambda)$.

I should note that we interpret $0! = 1$, so

$$p(0) = \mathrm{e}^{-\lambda}\frac{\lambda^0}{0!} = \mathrm{e}^{-\lambda}\frac{1}{1} = \mathrm{e}^{-\lambda}.$$



The Poisson distribution is named after the French mathematician Siméon-Denis Poisson who wrote about it in 1837, although the origin of the idea is more than 100 years earlier with another French mathematician, Abraham de Moivre.

Example 6.3. *An insurance company receives large insurance claims of over £100,000 at a rate of $\lambda = 1.2$ per day, modelled as a Poisson distribution and independent between days. What's the probability that in a week (5 days) they get at least one large claim every day?*

Let $X \sim \text{Po}(\lambda)$ be the number of large claims received in a day. Then the probability there is at least one claim in a day – which is an "at least one" question, suggesting we look at the complement – is

$$\mathbb{P}(X \geq 1) = 1 - \mathbb{P}(X = 0) = 1 - \mathrm{e}^{-1.2}\frac{1.2^0}{0!} = 1 - \mathrm{e}^{-1.2} = 0.699.$$

Since it's assumed that days are independent, the probability there is at least one large claim all 5 days is

$$\mathbb{P}(X \geq 1)^5 = 0.699^5 = 0.167,$$

or about 17%.

The parameter λ is called the "rate" because that indeed the number of emails (or insurance claims, or phone calls, or deaths by horse-kicking) that we expect to see.

Theorem 6.3. *Let $X \sim Po(\lambda)$. Then*

- *$p(x)$ is indeed a PMF, in that $\displaystyle\sum_{x=0}^{\infty} p(x) = 1$.*

- $\mathbb{E}X = \lambda$,
- $\mathrm{Var}(X) = \lambda$.

Proof. We'll do the first two here, then you can do the variance in Problem Sheet 4.

It will be useful to remember the Taylor series for the exponential function,

$$\mathrm{e}^{\lambda} = \sum_{x=0}^{\infty} \frac{\lambda^k}{x!}.$$

To see that the PMF does indeed sum to one, note that the Taylor series gives us

$$\sum_{x=0}^{\infty} p(x) = \sum_{x=0}^{\infty} \mathrm{e}^{-\lambda} \frac{\lambda^x}{x!} = \mathrm{e}^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = \mathrm{e}^{-\lambda}\, \mathrm{e}^{\lambda} = 1.$$

For the expectation, we have

$$\begin{aligned}
\mathbb{E}X &= \sum_{x=0}^{\infty} x\, \mathrm{e}^{-\lambda} \frac{\lambda^x}{x!} \\
&= \mathrm{e}^{-\lambda} \sum_{x=1}^{\infty} x\, \frac{\lambda^x}{x!} \\
&= \mathrm{e}^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^x}{(x-1)!} \\
&= \lambda \mathrm{e}^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!}
\end{aligned}$$

In the second line, we took $\mathrm{e}^{-\lambda}$ outside the sum, and allowed ourselves to start the sum from 1, since the $x = 0$ term was 0 anyway; in the third line, we cancelled the x from the $x!$ to get $(x-1)!$; and in the fourth line we took one of the λs in λ^x outside the sum, to give ourselves terms in $x-1$ inside the sum. We can now "re-index" the sum by putting $y = x - 1$, to get

$$\mathbb{E}X = \lambda \mathrm{e}^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} = \lambda \mathrm{e}^{-\lambda} \mathrm{e}^{\lambda} = \lambda,$$

where we used the Taylor series again. □

6.4 Poisson approximation to the binomial

Suppose I own a watch shop in Leeds. My watches are very expensive, so I don't need to sell many each day – in fact, I sell an average of 4.8 watches per day. How should I model the number of watches sold each day as a random variable?

One way could be to say this. There are n people living in Leeds or nearby, and, on any given day, each of them will independently buy a watch from my shop

with probability p. Thus the total number of watches I sell could be modelled as a binomial distribution $\mathrm{Bin}(n, p)$.

But what should n and p be? To make the average $\mathbb{E}X = np = 4.8$, I should take $p = 4.8/n$. But what about n? We know n is a very big number, because Leeds is a big city, so let's take a limit as $n \to \infty$. It turns out, that this distribution $\mathrm{Bin}(n, 4.8/n)$ becomes a Poisson(4.8) distribution!

Theorem 6.4. *Fix $\lambda \geq 0$, and let $X_n \sim Bin(n, \lambda/n)$ for all integers $n \geq \lambda$. Then $X_n \to Po(\lambda)$ in distribution as $n \to infty$, by which we mean that if $Y \sim Po(\lambda)$, then*

$$p_{X_n}(x) \to p_Y(x) \qquad \textit{for all } x \in \{0, 1, \dots\}.$$

A looser way to state the principle of this theorem would be this: *When n is very large and p very small, in such a way that np is a small-ish number, then $Bin(n, p)$ is well approximated by $Po(\lambda)$ where $\lambda = np$.*

This is why a Poisson distribution is a good model for the number of occurrences in a set time period. It applies if there lots of things that could happen (large n), each one is individually unlikely (small p), and on average a few of them will actually happen ($\lambda = np$ small-ish).

Example 6.4. *A lecturer teaches a module with $n = 100$, and estimate that each student turns up to office hours drop-in sessions independently with probability $p = 0.035$. What is the probability that **(a)** exactly 5, **(b)** 2 or more students turn up to a drop-in session?*

If we let X be the number of students that turn up to a drop-in session, then the exact distribution of X is $X \sim \mathrm{Bin}(100, 0.035)$.

For part (a), we then have

$$\mathbb{P}(X = 5) = \binom{100}{5} 0.035^5 (1 - 0.035)^{100-5} = 0.134.$$

For part (b), we use the complement rule to get

$$\begin{aligned}
\mathbb{P}(X \geq 2) &= 1 - \mathbb{P}(X = 0) - \mathbb{P}(X = 1) \\
&= 1 - \binom{100}{0} 0.035^0 (1 - 0.035)^{100-0} + \binom{100}{1} 0.035^1 (1 - 0.035)^{100-1} \\
&= 1 - (1 - 0.035)^{100} + 100 \times 0.035 (1 - 0.035)^{99} \\
&= 1 - 0.028 - 0.103 \\
&= 0.869
\end{aligned}$$

Alternatively, it might be more convenient to approximate X by a Poisson distribution $Y \sim \mathrm{Po}(100 \times 0.035) = \mathrm{Po}(3.5)$.

For part (a), this gives

$$\mathbb{P}(Y = 5) = \mathrm{e}^{-3.5} \frac{3.5^5}{5!} = 0.132,$$

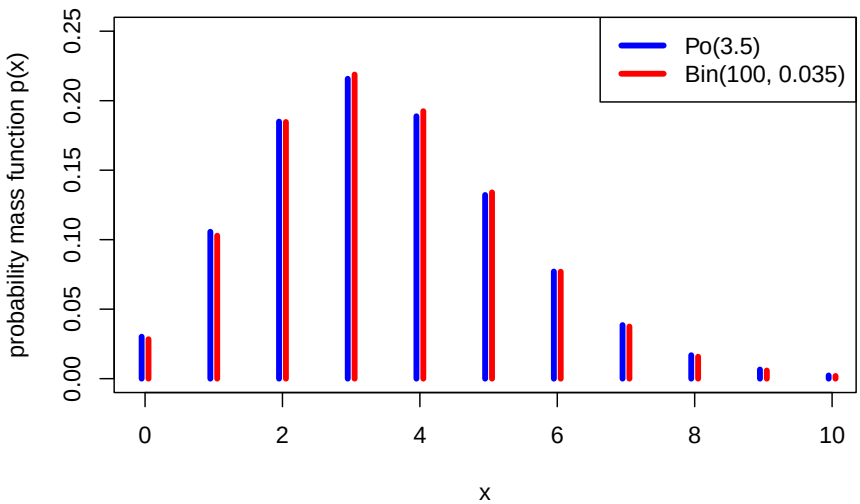
which is very close to the exact answer above of 0.134.

For part (b), the approximation gives

$$\begin{aligned}
\mathbb{P}(Y \geq 2) &= 1 - \mathbb{P}(Y = 0) - \mathbb{P}(Y = 1) \\
&= 1 - \mathrm{e}^{-3.5}\frac{3.5^0}{0!} - \mathrm{e}^{-3.5}\frac{3.5^1}{1!} \\
&= 1 - \mathrm{e}^{-3.5} - 3.5\mathrm{e}^{-3.5} \\
&= 1 - 0.030 - 0.106 \\
&= 0.864
\end{aligned}$$

which is very close to the exact answer above of 0.869.

The following graph shows how close the Po(3.5) distribution is to a Bin(100, 0.035) distribution – not exact, but pretty good.



For completeness, we include a proof of Theorem 6.4 here, although since it discusses use of limits, it's not examinable material for this module.

Proof. (Non-examinable) We need to show that, as $n \to \infty$,

$$\binom{n}{x}\left(\frac{\lambda}{n}\right)^x\left(1 - \frac{\lambda}{n}\right)^{n-x} \to \mathrm{e}^{-\lambda}\frac{\lambda^x}{x!}.$$

Let's try! The left-hand side is

$$\begin{aligned}
&\binom{n}{x}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x}\\
&\quad=\frac{n(n-1)\cdots(n-x+1)}{x!}\frac{\lambda^x}{n^x}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-x}\\
&\quad=\frac{\lambda^x}{x!}\frac{n(n-1)\cdots(n-x+1)}{n^x}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-x}\\
&\quad=\frac{\lambda^x}{x!}\frac{n}{n}\frac{n-1}{n}\cdots\frac{n-x+1}{n}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-x}\\
&\quad=\frac{\lambda^x}{x!}1\left(1-\frac{1}{n}\right)\cdots\left(1-\frac{x-1}{n}\right)\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-x}.
\end{aligned}$$

We take each of the terms in turn. First $\lambda^x/x!$ looks very promising, and can stay. Second, each of the terms $1, 1-1/n, \ldots, 1-(x-1)/n$ tend to 1 as $n \to \infty$. Third,

$$\left(1-\frac{\lambda}{n}\right)^n \to \mathrm{e}^{-\lambda};$$

this is from a standard result that

$$\left(1+\frac{a}{n}\right)^n \to \mathrm{e}^a \qquad \text{as } n \to \infty.$$

Finally

$$\left(1-\frac{\lambda}{n}\right)^{-x} \to 1,$$

as $1-\lambda/n \to 1$, and x is fixed. Putting all that together gives the result. □

6.5 Distributions as models for data

Before finishing this section, I want to say a bit more about using these families of distributions as models for data. This is a preview of the Bayesian statistics we will discuss in Section 10 of this module, and previews an idea that is extremely important throughout the whole MATH1712 Probability and Statistics II.

The families of distributions we have looked at here are sometimes called "parametric families", in that each of the distributions depended on one or more parameters: p for the Bernoulli and geometric distributions; λ for the Poisson distribution; and both n and p for the binomial distribution. (Later in the module we will also see some continuous parametric families: the exponential, normal and beta distributions.) This means we can adopt a model that data comes from a family, then use the data to estimate that parameter.

For example:

- You might assume that the number of emails you receive each day Poisson distributed with some unknown rate λ. You can then collect some data about how many emails you received each day for some large number of days, and use this data to estimate the value of λ.

- A tutor might assume that the number of students that turn up to each tutorial is binomially distributed where n is known to be 12, but p is unknown. The tutor could then take records of how many students turned up to all the tutorials, and use this to estimate p.

There are two main methods statisticians use to estimate parameters:

- **Frequentist statistics:** Here, one seeks a parameter that makes the data that was actually seen very likely. For example, if the mean number of emails you received per day is $\bar{x}$, then this suggests that taking $\lambda = \bar{x}$, which leads to the expected number of emails being $\bar{x}$ also, could be a good estimate. The frequentist approach will pursued in detail in MATH1712.
- **Bayesian statistics:** Here, one starts with a "prior" distribution for the parameters, which represents one's belief about which possible values for that parameter are more or less likely. One then uses Bayes' theorem to update that belief to a "posterior" distribution of one's beliefs about the parameter *given* the data. The Bayesian approach will be introduced in Section 10 of these notes, and there will be the opportunity to pursue this further later in your university career.

Summary

Distribution	Range	PMF	Expectation	Variance
Bernoulli: Bern(p)	$\{0, 1\}$	$p(0) = 1-p$, $p(1) = p$	p	$p(1-p)$
Binomial: Bin(n, p)	$\{0, 1, \dots, n\}$	$\binom{n}{x} p^x (1-p)^{n-x}$	np	$np(1-p)$
Geometric: Geom(p)	$\{1, 2, \dots\}$	$(1-p)^{x-1}p$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson: Po(λ)	$\{0, 1, \dots\}$	$\mathrm{e}^{-\lambda}\frac{\lambda^x}{x!}$	λ	λ

Chapter 7

Multiple random variables

7.1 Joint distributions

In previous sections, we have looked at single discrete random variables in isolation. In the section, we want to look at how multiple discrete random variables can interact.

Consider tossing a fair coin 3 times. Let X be the number of Heads in the first two tosses, and let Y be the number of Heads over all three tosses.

We know that $X \sim \text{Bin}(2, \frac{1}{2})$ and $Y \sim \text{Bin}(3, \frac{1}{2})$, so we can easily write down their probability mass functions:

x	$x = 0$	$x = 1$	$x = 2$
$p_X(x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

y	$y = 0$	$y = 1$	$y = 2$	$y = 3$
$p_Y(y)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

When we have multiple random variables and we want to emphasise that a PMF refers to only one of them, we often use the phrase **marginal PMF** or **marginal distribution**. So the PMFs above are the marginal distributions of X and Y.

However, we might also want to know how X and Y interact. To do this, we will need the **joint PMF**, given by

$$p_{X,Y}(x, y) = \mathbb{P}(X = x \text{ and } Y = y).$$

In our case of the coin tosses, we have

$p_{X,Y}(x,y)$	$y=0$	$y=1$	$y=2$	$y=3$
$x=0$	$\frac{1}{8}$	$\frac{1}{8}$	0	0
$x=1$	0	$\frac{1}{4}$	$\frac{1}{4}$	0
$x=2$	0	0	$\frac{1}{8}$	$\frac{1}{8}$

For probabilities of events, we had that if some A_is form a partition, then

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

Note that the events $\{X = x\}$, as x varies over the range of X, also make up a partition. Therefore, we have

$$\mathbb{P}(Y = y) = \sum_x \mathbb{P}(X = x \text{ and } Y = y);$$

or, to phrase this in terms of joint and marginal PMFs,

$$p_Y(y) = \sum_x p_{X,Y}(x,y).$$

In other words, to get the marginal distribution of Y, we need to sum down the columns in the table of the joint distribution.

$p_{X,Y}(x,y)$	$y=0$	$y=1$	$y=2$	$y=3$
$x=0$	$\frac{1}{8}$	$\frac{1}{8}$	0	0
$x=1$	0	$\frac{1}{4}$	$\frac{1}{4}$	0
$x=2$	0	0	$\frac{1}{8}$	$\frac{1}{8}$
$p_Y(y)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

In exactly the same way, we have

$$p_X(x) = \sum_y p_{X,Y}(x,y);$$

so to get the marginal distribution of X, we need to sum across the rows in the table of the joint distribution.

$p_{X,Y}(x,y)$	$y=0$	$y=1$	$y=2$	$y=3$	$p_X(x)$
$x=0$	$\frac{1}{8}$	$\frac{1}{8}$	0	0	$\frac{1}{4}$
$x=1$	0	$\frac{1}{4}$	$\frac{1}{4}$	0	$\frac{1}{2}$
$x=2$	0	0	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$
$p_Y(y)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	

We can check that these marginal PMFs match those we started with. The term "marginal" PMF or distribution is presumably because one ends up writing the values in the "margins" of the table.

Note that the joint PMF conforms to the same rules as a normal PMF:

- it is non-negative: $p_{X,Y}(x,y) \geq 0$;
- it sums to 1: $\sum_{x,y} p_{X,Y}(x,y) = 1$.

We may want to look at more than two random variables, $\mathbf{X} = (X_1, X_2, \dots, X_n)$. In this case, the joint PMF is

$$p_{\mathbf{X}}(\mathbf{x}) = p_{X_1,\dots,X_n}(x_1, \dots, x_n) = \mathbb{P}(X_1 = x_1 \text{ and } \cdots \text{ and } X_n = x_n).$$

In the same way, we can find the marginal distribution of one of the variables – say X_1 – by summing over all the other variables:

$$p_{X_1}(x_1) = \sum_{x_2,\dots,x_n} p_{X_1,X_2,\dots,X_n}(x_1, x_2, \dots, x_n).$$

7.2 Independence of random variables

We said that two events are independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$. We now can give a similar definition for what it means two random variables to be independent.

Definition 7.1. We say two discrete random variables are **independent** if, for all x and y,

$$\mathbb{P}(X = x \text{ and } Y = y) = \mathbb{P}(X = x)\,\mathbb{P}(Y = y).$$

In terms of the joint and marginal PMFs, this is the condition that

$$p_{X,Y}(x,y) = p_X(x)\,p_Y(y).$$

More generally, a sequence of random variable $X_1, X_2, \dots$ are independent if

$$p_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} p_{X_i}(x_i).$$

Returning to our case of the dice from before, we see that X and Y are not independent, because, for just one counterexample, $p_{X,Y}(0,0) = \frac{1}{8}$, while $p_X(0) = \frac{1}{4}$ and $p_Y(0) = \frac{1}{8}$.

An important scenario in probability theory and statistics is that of **independent and identically distributed** (or **IID**) random variables. IID random variables represent an experiment that is repeated many times, with each experiment independent of the others. So all the random variables have the same distribution, and they are all independent of each other. So $\mathbf{X} = (X_1, X_2, \dots, X_n)$ are IID random variables with a common PMF p_X, say, if

$$p_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} p_X(x_i).$$

Example 7.1. *Let $X_1, X_2, \dots, X_{10}$ be IID random variables following a Poisson distribution with rate $\lambda = 3$. What is the probability that all the X_i are nonzero?*

Because the X_i are identically distribution, the probability that any one of them is nonzero is

$$\mathbb{P}(X_1 \geq 0) = 1 - \mathbb{P}(X_1 = 0) = 1 - \mathrm{e}^{-3} = 0.950.$$

Then, because the X_i independent, the probability that they are all nonzero is

$$\mathbb{P}(X_1 \geq 0)^2 0 = 0.950^2 0 = 0.360.$$

7.3 Conditional distributions

For probability with events, we had the conditional probability

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

In the same way, for random variables, we have

$$\mathbb{P}(Y = y \mid X = x) = \frac{\mathbb{P}(X = x \text{ and } Y = y)}{\mathbb{P}(X = x)}.$$

It makes sense to call the distribution of this the **conditional distribution**.

Definition 7.2. Let X and Y be two random variables with joint PMF $p_{X,Y}$ and marginal PMFs p_X and p_Y respectively. Then the **condition probability mass function of Y given X** $p_{Y|X}$ is given by

$$p_{Y|X}(y \mid x) = \frac{p_{X,Y}(x, y)}{p_X(x)}.$$

Let's think again about our coin tossing example. To get the conditional distribution of Y given $X = 1$, say, we have

$$p_{Y|X}(y \mid 1) = \frac{p_{X,Y}(1, y)}{p_X(1)};$$

so we take the $x = 1$ row of the joint distribution table and "renormalise it" by dividing through by the total of the row, so it adds up to 1. That is,

$$\begin{aligned}
p_{Y|X}(0 \mid 1) &= \frac{p_{X,Y}(1,0)}{p_X(1)} = \frac{0}{\frac{1}{2}} = 0, \\
p_{Y|X}(1 \mid 1) &= \frac{p_{X,Y}(1,1)}{p_X(1)} = \frac{\frac{1}{4}}{\frac{1}{2}} = \tfrac{1}{2}, \\
p_{Y|X}(2 \mid 1) &= \frac{p_{X,Y}(1,2)}{p_X(1)} = \frac{\frac{1}{4}}{\frac{1}{2}} = \tfrac{1}{2}, \\
p_{Y|X}(3 \mid 1) &= \frac{p_{X,Y}(1,3)}{p_X(1)} = \frac{0}{\frac{1}{2}} = 0.
\end{aligned}$$

In just the same way, we could get the conditional distribution of X given $Y = 2$, say, by taking the $y = 2$ column of the joint distribution table, and renormalising so that the column sums to 1, That is,

$$p_{X|Y}(0 \mid 2) = \frac{p_{X,Y}(0,2)}{p_Y(2)} = \frac{0}{\frac{3}{8}} = 0,$$
$$p_{X|Y}(1 \mid 2) = \frac{p_{X,Y}(1,2)}{p_Y(2)} = \frac{\frac{1}{4}}{\frac{3}{8}} = \tfrac{2}{3},$$
$$p_{X|Y}(2 \mid 2) = \frac{p_{X,Y}(2,2)}{p_Y(2)} = \frac{\frac{1}{8}}{\frac{3}{8}} = \tfrac{1}{3}.$$

Results that we used for conditional probability with events also carry over to random variables. For example, from **Bayes' theorem** we know that

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

In the same way, we have

$$\mathbb{P}(X = x \mid Y = y) = \frac{\mathbb{P}(X = x)\,\mathbb{P}(Y = y \mid X = x)}{\mathbb{P}(Y = y)},$$

which in terms of conditional and marginal PMFs is

$$p_{X|Y}(x \mid y) = \frac{p_X(x)\,p_{Y|X}(y \mid x)}{p_Y(y)}.$$

This will be a particularly important formula when we study Bayesian statistics at the end of the module.

We can check Bayes' theorem with $x = 1$ and $y = 2$, for example. The right-hand side of Bayes' theorem is

$$\frac{p_X(1)\,p_{Y|X}(2 \mid 1)}{p_Y(2)} = \frac{\frac{1}{2} \times \frac{1}{2}}{\frac{3}{8}} = \frac{23}{.}$$

The left-hand side of Bayes' theorem is

$$p_{X|Y}(1 \mid 2) = \tfrac{2}{3},$$

which is equal, as it should be, to the right-hand side.

7.4 Expectation sums and products

When we have multiple random variables, we might be interested in functions of those multiple random variables – for example their sum or their product. (In fact, it's often possible to find out about the whole distribution of a sum, product, or function of the variables – see MATH2715 Statistical Methods for more on this – but will just look at their expectations and, later, variances.)

Theorem 7.1. *Let X and Y be two random variables with joint probability mass function $p_{X,Y}$. Then*

1. $\mathbb{E}g(X,Y) = \sum_{x,y} g(x,y)p_{X,Y}(x,y)$.
2. ***(Linearity of expectation, 2)*** $\mathbb{E}(X+Y) = \mathbb{E}X + \mathbb{E}Y$, *regardless of whether X and Y are independent or not.*
3. *If X and Y are independent, then $\mathbb{E}XY = \mathbb{E}X \times \mathbb{E}Y$.*

If we put the second point here together with the other result of linearity of expectation (Theorem 5.3) then we get the general rule

$$\mathbb{E}(aX + bY + c) = a\,\mathbb{E}X + b\,\mathbb{E}Y + c,$$

and this holds whether or not X and Y are independent.

Proof. Part 1 is just the law of the unconscious statistician for the random variable (X,Y), and the same proof holds.

For part 2, we have

$$\begin{aligned}
\mathbb{E}(X+Y) &= \sum_{x,y}(x+y)p_{X,Y}(x,y) \\
&= \sum_{x,y} x\,p_{X,Y}(x,y) + \sum_{x,y} y\,p_{X,Y}(x,y) \\
&= \sum_{x} x \sum_{y} p_{X,Y}(x,y) + \sum_{y} y \sum_{x} p_{X,Y}(x,y)
\end{aligned}$$

But summing a joint PMF over one of the variables gives the marginal PMF; so $\sum_y p_{X,Y}(x,y) = p_X(x)$ and $\sum_x p_{X,Y}(x,y) = p_Y(y)$. So this gives

$$\begin{aligned}
\mathbb{E}(X+Y) &= \sum_{x} x\,p_X(x) + \sum_{y} y\,p_Y(y) \\
&= \mathbb{E}X + \mathbb{E}Y.
\end{aligned}$$

For part 3, if X and Y are independent, then $p_{X,Y}(x,y) = p_X(x)\,p_Y(y)$. Therefore,

$$\begin{aligned}
\mathbb{E}XY &= \sum_{x,y} xyp_{X,Y}(x,y) \\
&= \sum_{x}\sum_{y} xyp_X(x)p_Y(y) \\
&= \sum_{x} xp_X(x) \sum_{y} yp_Y(y) \\
&= \mathbb{E}X \times \mathbb{E}Y,
\end{aligned}$$

as required. □

Example 7.2. *A student is solving five questions on a problem sheet. The time taken for each question to the nearest minute is an identically distributed random variable with expectation $\mathbb{E}X_i = \mu$. What is the total expected time to complete the problem sheet?*

By linearity of expectation, this is

$$\mathbb{E}(X_1 + X_2 + X_3 + X_4 + X_5) = \mathbb{E}X_1 + \mathbb{E}X_2 + \mathbb{E}X_3 + \mathbb{E}X_4 + \mathbb{E}X_5 = 5\mu.$$

What if the lengths of time are not independent – for example, if the student is slower at answering all the questions when she is tired?

It's still the case that $\mathbb{E}(X_1 + X_2 + X_3 + X_4 + X_5) = 5\mu$, because this result is true whether or not the random variables are independent.

7.5 Covariance

If we are interested at how two random variables vary together, we need to look at the covariance.

Definition 7.3. Let X and Y be two random variables with expectations μ_X and μ_Y respectively. Then their **covariance** is

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

In the least surprising result of this whole module, we also have a computational formula to go along with this definitional formula.

Theorem 7.2. *Let X and Y be two random variables with expectations μ_X and μ_Y respectively. Then their covariance can also be calculated as*

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

Proof. Exactly as we've done many times before, we have

$$\begin{aligned}
\mathrm{Cov}(X, Y) &= \mathbb{E}(X - \mu_X)(Y - \mu_Y) \\
&= \mathbb{E}(XY - X\mu_Y - \mu_X Y + \mu_X \mu_Y) \\
&= \mathbb{E}XY - \mu_Y \mathbb{E}X - \mu_X \mathbb{E}Y + \mu_X \mu_Y \\
&= \mathbb{E}XY - \mu_X \mu_Y - \mu_X \mu_Y + \mu_X \mu_Y \\
&= \mathbb{E}XY - \mu_X \mu_Y,
\end{aligned}$$

and we're done. □

A very important fact is the following.

Theorem 7.3. *If X and Y are independent, then* $\mathrm{Cov}(X, Y) = 0$.

Proof. Recall from Theorem 7.1 that if X and Y are independent, we have $\mathbb{E}XY = \mathbb{E}X \times \mathbb{E}Y = \mu_X \mu_Y$. Then from the computational formula, we have

$$\mathrm{Cov}(X, Y) = \mathbb{E}XY - \mu_X \mu_Y = \mu_X \mu_Y - \mu_X \mu_Y = 0,$$

and we are done. □

Here are some more important properties of the covariance.

Theorem 7.4. *Let X, Y and Z be random variables. Then*

1. $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$;
2. $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$'
3. $\mathrm{Cov}(aX, Y) = a\ \mathrm{Cov}(X, Y)$;
4. $\mathrm{Cov}(X + b, Y) = \mathrm{Cov}(X, Y)$;
5. $\mathrm{Cov}(X + Y, Z) = \mathrm{Cov}(X, Z) + \mathrm{Cov}(Y, Z)$.

Proof. Part 1 and 2 are immediate from the definition.

Parts 3, 4 and 5 are quite similar. We'll do part 5 here, and you can do parts 3 and 4 on Problem Sheet 4.

For part 5, note that $\mathbb{E}(X + Y) = \mu_X + \mu_Y$ by linearity of expectation. Hence

$$\begin{aligned}
\mathrm{Cov}(X + Y, Z) &= \mathbb{E}(X + Y - \mu_X - \mu_Y)(Z - \mu_Z) \\
&= \mathbb{E}\big((X - \mu_X) + (Y - \mu_Y)\big)(Z - \mu_Z) \\
&= \mathbb{E}\big((X - \mu_X)(Z - \mu_Z) + (Y - \mu_Y)(Z - \mu_Z)\big) \\
&= \mathbb{E}(X - \mu_X)(Z - \mu_Z) + \mathbb{E}(Y - \mu_Y)(Z - \mu_Z) \\
&= \mathrm{Cov}(X, Z) + \mathrm{Cov}(Y, Z),
\end{aligned}$$

as required. □

Now that we know some facts about the covariance, we can calculate the variance of a sum.

Theorem 7.5. *Let X and Y be two random variables. Then*

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + 2\,\mathrm{Cov}(X, Y) + \mathrm{Var}(Y).$$

If X and Y are independent, then

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

It's easy to forget the conditions for the following two facts:

- $\mathbb{E}(X + Y) = \mathbb{E}X + \mathbb{E}Y$ regardless of whether X and Y are independent or not.
- $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ if X and Y are independent.

Proof. For the main part of the proof, we start with the definition of variance. By linearity of expectation, we have $\mathbb{E}(X + Y) = \mu_X + \mu_Y$. So

$$\begin{aligned}
\mathrm{Var}(X + Y) &= \mathbb{E}\big((X + Y) - (\mu_X + \mu_Y)\big)^2 \\
&= \mathbb{E}\big((X - \mu_X) + (Y - \mu_Y)\big)^2 \\
&= \mathbb{E}\big((X - \mu_X)^2 + 2(X - \mu_X)(Y - \mu_Y) + (Y - \mu_Y)^2\big) \\
&= \mathbb{E}(X - \mu_X)^2 + 2\mathbb{E}(X - \mu_X)(Y - \mu_Y) + \mathbb{E}(Y - \mu_Y)^2 \\
&= \mathrm{Var}(X) + 2\,\mathrm{Cov}(X, Y) + \mathrm{Var}(Y),
\end{aligned}$$

where we used the linearity of expectation.

For the second part, recall that is X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$. $\square$

Definition 7.4. Let X and Y be two random variables. Then the **correlation** between X and Y is

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

As with the sample correlation r_{xy} from Section 1, the correlation is a number between -1 and $+1$, where values near $+1$ mean that large values of X and large values of Y are likely to occur together, while values near -1 mean that large values of X and small values of Y are likely to occur together.

Recall that, if X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$. Hence it follows that if X and Y are idnependent, then $\mathrm{Corr}(X, Y) = 0$ also.

7.6 Law of large numbers

Let $X_1, X_2, \dots, X_n$ be a sequence of IID random variables. Let us write $\mu = \mathbb{E}X_1$ for the common expectation and $\sigma^2 = \mathrm{Var}(X_1)$ for the common variance.

At the beginning of the course, we saw the mean of some values was

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i;$$

that is, what we get if we add them up and divide by n. In the same way, we could calculate the "mean" of some random variables by adding them up and dividing by n; that is:

$$\overline{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

(The subscript n on "$\overline{X}_n$" is just to remind us this is a mean of n random variables.)

Here, each of the X_is is a random variable, so their mean $\overline{X}_n$ is another random variable as well. So we can ask questions about the random variable $\overline{X}_n$ just the same as we would ask about any other random variable. For example: What is its expectation and variance?

The expectation of $\bar{X}_n$ is

$$\begin{aligned}
\mathbb{E}\overline{X}_n &= \mathbb{E}\left(\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right) \\
&= \frac{1}{n}(\mathbb{E}X_1 + \mathbb{E}X_2 + \cdots + \mathbb{E}X_n) \\
&= \frac{1}{n}(\mu + \mu + \cdots + \mu) \\
&= \frac{1}{n}n\mu \\
&= \mu.
\end{aligned}$$

Here we use linearity of expectation to take the $1/n$ out of the brackets, and then add up the individual expectations.

Since the X_is are independent, the variance of $\overline{X}_n$ is

$$\begin{aligned}
\mathrm{Var}(\overline{X}_n) &= \mathrm{Var}\left(\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right) \\
&= \left(\frac{1}{n}\right)^2 \mathrm{Var}(X_1 + X_2 + \cdots + X_n) \\
&= \frac{1}{n^2}(\mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathbb{E}X_n) \\
&= \frac{1}{n^2}(\sigma^2 + \sigma^2 + \cdots + \sigma^2) \\
&= \frac{1}{n^2}n\sigma^2 \\
&= \frac{\sigma^2}{m}.
\end{aligned}$$

In conclusion we have this:

Theorem 7.6. *Let $X_1, X_2, \ldots, X_n$ be a sequence of IID random variables. Write $\mu = \mathbb{E}X_1$ for the common expectation and $\sigma^2 = \mathrm{Var}(X_1)$ for the common variance, and $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ for the mean. Then*

$$\mathbb{E}\overline{X}_n = \mu \qquad \mathrm{Var}(\overline{X}_n) = \frac{\sigma^2}{n}.$$

Now think about what happens to this mean $\overline{X}_n$ when n gets very large. We see that the expectation $\mathbb{E}\overline{X}_n = \mu$ stays the same, but the variance $\mathrm{Var}(\overline{X}_n) = \sigma^2/n$ gets smaller and smaller as n gets bigger. Thus the range of probably values for $\overline{X}_n$ well be squeezing tighter and tighter around μ. Given that, it seems as if (and it can be proven that) we have the "law of large numbers".

Theorem 7.7 (Law of large numbers)**.** *Let $X_1, X_2, \ldots$ be a sequence of IID random variables. Write $\mu = \mathbb{E}X_1$ for the common expectation and $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ for the mean of the first n random variables. Then*

$$\overline{X}_n \to \mu \quad \textit{in distribution as } n \to \infty;$$

by which we mean that, for any $\epsilon > 0$,

$$\mathbb{P}(|\overline{X}_n - \mu| > \epsilon) \to 0 \quad \textit{as } n \to \infty.$$

The precise mathematical definition of the convergence is not important here. What is important is the general principle that the expectation represents the "long-run average" of independent experiments.

One special case is if we have repeated experiments the succeed with probability p; that is, $X_n \sim \text{Bern}(p)$. Then the law of large numbers says that the long-run proportion of successes is

$$\frac{1}{n}\sum_{i=1}^{n} X_n = \overline{X}_n \to \mathbb{E}X_1 = p.$$

So the long-run proportion of times an event happens converges to its probability. This goes right back to what we said about "frequentist probability" right at the beginning of Section 2: that one way to understand the probability of an event is as the long-run frequency of its occurrence.

Problem Sheet 4

This is Problem Sheet 4. This problem sheet covers Section 6, and next week will have questions added on Section 7. You should work through all the questions on this problem sheet so far during Week 6, in preparation for your tutorial in Week 8. The problem sheet contain an assessed question, to which a second will be added next week, and which are due in by **2pm on Monday 22 November**.

Please also fill in the mid-semester survey if you haven't done so yet.

A: Short questions

A1. Let $X \sim \text{Bin}(20, 0.4)$. Calculate

(a) $\mathbb{P}(X = 8)$

(b) $\mathbb{P}(8 \leq X \leq 11)$

(c) $\mathbb{E}X$

A2. Let $X \sim \text{Geom}(0.2)$. Calculate

(a) $\mathbb{P}(X = 2)$

(b) $\mathbb{P}(X \geq 3)$

(c) $\text{Var}(X)$

A3. Let $X \sim \text{Po}(2.5)$. Calculate

(a) $\mathbb{P}(X = 3)$

(b) $\mathbb{P}(X \geq \mathbb{E}X)$

A4. Consider the following joint PMF:

$p_{X,Y}(x, y)$	$y = 0$	$y = 1$	$y = 2$	$y = 3$
$x = 0$	$2k$	$2k$	k	0
$x = 1$	k	$3k$	k	k
$x = 2$	0	k	k	$2k$

(a) Find the value of k that makes this a joint PMF.

(b) Find the marginal PMFs of X and Y.

(c) What is the conditional distribution of Y given $X = 1$?

(d) Are X and Y independent?

A5. Let X and Y be Bernoulli($\frac{1}{2}$) random variables.

(a) Write down the table for the joint PMF of X and Y if X and Y are independent.

(b) Write down a table for a joint PMF of X and Y that is consistent with their marginal distributions but that leads to X and Y having a positive correlation.

(c) Write down a table for a joint PMF of X and Y that is consistent with their marginal distributions but that leads to X and Y having a negative correlation.

B: Long questions

B1. Calculate the CDF $F(x) = \mathbb{P}(X \leq x)$ of the geometric distribution…

(a) …by summing the PMF;

(b) …by explaining how the "number of trials until success" definition tells us what $1 - F(x) = \mathbb{P}(X > x)$ must be.

(c) A gambler rolls a pair of dice until he gets a double-six. What is the probability that this takes between 20 and 40 double-rolls?

B2. Let Y be a geometric distribution with parameter p according to the alternative "number of failures *before* the first success" definition.

(a) Write down the PMF for Y.

(b) Calculate the expectation and variance of Y. You may use without proof the fact that for a standard "number of trials up to and including the first success" geometric distribution we have $\mathbb{E}X = 1/p$ and $\mathrm{Var}(X) = (1-p)/p^2$.

B3 Let $X \sim \mathrm{Po}(\lambda)$.

(a) Show that $\mathbb{E}X(X-1) = \lambda^2$. You may use the Taylor series for the exponential,

$$\mathrm{e}^{\lambda} = \sum_{y=0}^{\infty} \frac{\lambda^y}{y!}.$$

(b) Hence show that $\mathrm{Var}(X) = \lambda$. You may use the fact, proved in the notes, that $\mathbb{E}X = \lambda$.

B4. Each week in the UK about 15 million Lotto tickets are sold. As we saw in Section 3, the probability of each ticket winning is about 1 in 45 million. Estimate the proportion of weeks when there is **(a)** a roll-over (no jackpot winners), **(b)** a unique jackpot winner, or **(c)** when multiple winners share the jackpot. State any modelling assumptions you make and the approximation that you use.

B5. Let $X_1, X_2, \dots, X_n$ be IID random variable with common expectation μ and common variance σ^2, and let $\overline{X} = (X_1 + \cdots + X_n)n$ be the mean of these random variables. We will be considering the random variable S^2 given by

$$S^2 = \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

(a) By writing

$$X_i - \overline{X} = (X_i - \mu) - (\overline{X} - \mu)$$

or otherwise, show that

$$S^2 = \sum_{i=1}^{n} (X_i - \mu)^2 - n(\overline{X} - \mu)^2.$$

(b) Hence or otherwise, show that

$$\mathbb{E}S^2 = (n-1)\sigma^2.$$

You may use facts about $\overline{X}$ from the notes provided you state them clearly. You might also find it useful to recall that if a random variable Y has expectation $\mathbb{E}Y = 0$, then $\mathbb{E}Y^2 = \mathrm{Var}(Y)$.

(c) At the beginning of this course, we defined the sample variance of the values $x_1, x_2, \dots, x_n$ to be

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Explain one reason why we might consider it appropriate to use $1/(n-1)$ as the factor at the beginning of this expression, rather than simply $1/n$.

C: Assessed questions

The last two questions are **assessed questions**. These two questions count for 3% of your final mark for this module.

The deadline for submitting your solutions is **2pm on Monday 22 November** at the beginning of Week 9, although I strongly recommend completing and submitting your work during Week 8. Submission will be via Gradescope; submission will open on Monday 15 November. Your work will be marked by your tutor and returned on Monday 29 November, when solutions will also be made available.

Both questions are "long questions", where the marks are not only for mathematical accuracy but also for the clarity and completeness of your explanations.

You should not collaborate with others on the assessed questions: your answers must represent solely your own work. The University's rules on academic integrity – and the related punishments for violating them – apply to your work on the assessed questions.

C1. A collector wants to collect football stickers to fill an album. There are n unique stickers to collect. Each time the collector buys a sticker, it is one of the n stickers chosen independently uniformly at random. Unfortunately, it is likely the collector will end up having "swaps", where he has received the same sticker more than once, so he will likely need to buy more than n stickers in total to fill his album. But how many?

(a) Suppose the collector has already got j unique stickers (and some number of swaps). Let X_j be the the number of extra stickers he buys until getting a new unique sticker. Explain why X_j is geometrically distributed, and state the parameter $p = p_j$ of the geometric distribution.

(b) Hence, show that the expected number of stickers the collector must buy to fill his album is

$$n \sum_{k=1}^{n} \frac{1}{k}.$$

(c) The Euro 2020 sticker album required $n = 678$ unique stickers to complete it, and stickers cost 15p each. Using the expression from (b), calculate the amount of money needed to fill the album. You should do this calculation in R and include the command you used in your answer.

(d) By approximating the sum in part (b) by an integral, explain why the expected number of stickers required is approximately $n \log n$, where log denotes the natural logarithm to base e.

C2. Let X and Y be random variables, and let a and b be constants.

(a) Starting from the definition of covariance, show that $\mathrm{Cov}(aX, Y) = a\,\mathrm{Cov}(X, Y)$. You may find it helpful to remember that if $\mathbb{E}X = \mu_X$, then $\mathbb{E}aX = a\mu_X$.

(b) Show that $\mathrm{Cov}(X + b, Y) = \mathrm{Cov}(X, Y)$.

Now let X, Y, Z be *independent* random variables with common variance σ^2.

(c) Find the value of $\mathrm{Corr}(2X - 3Y + 4, 2Y - Z - 1)$. You may use any facts about covariance from the notes, including those from parts (a) and (b) of this question, provided you state them clearly.

Solutions to short questions

A1. (a) 0.180 (b) 0.528 (c) 8 **A2.** (a) 0.16 (b) 0.64 (c) 20 **A3.** (a) 0.214 (b) 0.456

Other stuff

R Worksheets

R worksheets

Each week there will be an R worksheet to work through in your own time. We recommend spending about one hour on each worksheet, plus one extra hour for worksheets with assessed questions, for checking through and submitting your solutions.

Week	Worksheet	Deadline for assessed work
1	**R basics** (Solutions)	—
2	**Vectors**	—
3	**Data in R**	Monday 18 October
4	**Plots I:** Making plots	—
5	**Plots II:** Making plots better	Monday 1 November
6	**RMarkdown** (optional) [Rmd]	—
7	**Discrete distributions** [Rmd]	Monday 15 November
8	Discrete random variables	—
9	Normal distribution	Monday 29 November
10	Law of large numbers	—
11	Summary	*To be confirmed*

About R and RStudio

- **R** is a *programming language* that is particularly good at working with probability and statistics. R is very widely used in universities and increasingly widely used in industry. Learning to use R is a mandatory part of this module, and exercises requiring use of R make up at least 15% of your module mark. Many other statistics-related course at the University also use R.
- **RStudio** is a *program* that gives a convenient way to work with the language R. RStudio is the most common way to use the language R, and learning to use RStudio is strongly recommended.

R and RStudio are free/open-source software.

How to access R and RStudio

There are a number of ways you can access R and RStudio:

- All **University computers** have R and RStudio already installed. Here is a directory of the University's computer clusters.
- You can **install** R and RStudio on your own computer – see the instructions below.
- If you want to use R/RStudio on a non-University device for which you don't have admin/installation rights (Chromebook, iPad, friend's laptop, etc), you could try:
  - You can use the University's copies of R/RStudio virtually through the Windows Virtual Desktop or AppsAnywhere client.
  - The RStudio Cloud is a cloud-hosted "Google Docs for R" that you can use through your web browser – you can get 25 hours per month for free (or pay for more).

Installing R and RStudio

Students who have their own computer usually find it most convenient to install R and RStudio on that computer. To do this, it's important that you install R (the programming language) first, and only install RStudio (the program to use R) once R has already been installed.

1. *First*, install **R**. Go to the Comprehensive R Archive Network and follow the instructions:
   - Windows: Click "Download R for Windows", then "Install R for the first time". The main link at the top should be to download the most recent version of R.
   - Mac: Click Download R for macOS, and then download the relevant PKG file. (For pre-November 2020 Intel-based Macbooks, you must use the "Intel 64-bit build"; for post-November 2020 M1-based "Apple silicon" Macbooks, the "Apple silicon arm64 build" may be faster.)
2. *After* R is installed, *then* install **RStudio**. Go to the Download page at RStudio.com and follow the instructions. You want "RStudio Desktop", and you want the free version.

If you have difficulty installing R, come along to the first computational drop-in session in Week 2 and bring your computer with you (if it's sufficiently portable), and we'll do our best to help.

Troubleshooting drop-in sessions

You will learn to use R by working through the R Worksheets. Learning to use a programming language is different from learning mathematics: you should expect to regularly get frustrated and annoyed when the computer seems to refuse to do what you want it to (but also occasionally experience the joy of getting it right!). This is a normal part of learning.

However, many students find getting with started with R in the first few weeks particularly frustrating. Also, sometimes students have problems installing R and RStudio on their own computers. To help with this, we have organised optional troubleshooting drop-in sessions in Weeks 2 and 3. Check your timetable for details – they are probably listed as "computer practicals".

Solutions

This page has the solutions to all the non-assessed questions on Problem Sheet 1 to 3. Solutions are added on Friday afternoon after tutorials on the Problem Sheet have finished.

Solutions to assessed questions are available on Minerva in the "Assessments" tab, from one week after the deadline.

Problem Sheet 1

A1. Consider again the "number of Skittles in each packet" data from Example 1.1.

$$59,\ 59,\ 59,\ 59,\ 60,\ 60,\ 60,\ 61,\ 62,\ 62,\ 62,\ 63,\ 63.$$

(a) Calculate the mean number of Skittles in each packet.

Solution. This was in the notes:

$$\bar{x} = \frac{1}{13}(59 + 59 + \cdots + 63) = \frac{789}{13} = 60.7.$$

(b) Calculate the sample variance using the computational formula.

Solution.

$$\begin{aligned} s_x^2 &= \frac{1}{13-1}\left((59^2 + 59^2 + \cdots + 63^2) - 13 \times 60.6923^2)\right) \\ &= \frac{1}{12}(47915 - 47886.2) \\ &= 2.40 \end{aligned}$$

Group feedback: With the computational formula, the value $\sum_i x_i^2 - n\bar{x}^2$ is typically a fairly small number given as the difference between two very big numbers $\sum_i x_i^2$ and $n\bar{x}^2$. This means you have to get the two big numbers very precise, to ensure the cancellation happens correctly; in particular, make sure you use plenty of decimal places of accuracy in $\bar{x}$.

(c) Calculate the sample variance using the definitional formula.

Solution.

$$\begin{aligned} s_x^2 &= \frac{1}{13-1}\left((59-60.7)^2 + (59-60.7)^2 + \cdots + (63-60.7)^2\right) \\ &= \frac{1}{12}(2.86 + 2.86 + \cdots + 5.33) \\ &= \frac{1}{12} \times 28.77 \\ &= 2.40 \end{aligned}$$

(d) Out of (b) and (c), which calculation did you find easier, and why?

Solution. The computational formula required fewer presses of the calculator buttons, because $\sum_i x_i^2$ is fewer button-presses than $\sum_i (x_i - \bar{x})^2$, where you have to subtract the means before squaring.

On the other hand, the expression inside the brackets of the computational formula is a fairly small number given as the difference of two very large numbers, so it was necessary to use lots of decimal places of accuracy in $\bar{x}$ to make sure the second large number was accurate and therefore that the subtraction cancelled correctly.

A2. Consider the following data sets of the age of elected politicians on a local council. (The "18–30" consists of people older than and including 18, and younger than but *not* including 30.)

Age (years)	Frequency	Relative frequency	Frequency density
18–30	1		
30–40	3		
40–45	4		
45–50	5		
50–55	3		
55–60	1		
60–70	3		
Total	20	1	—

(a) Complete the table by filling in the relative frequency and frequency densities.

Solution.

Age (years)	Frequency	Relative frequency	Frequency density
18–30	1	0.05	0.0041
30–40	3	0.15	0.015
40–45	4	0.2	0.04
45–50	5	0.25	0.05
50–55	3	0.15	0.03
55–60	1	0.05	0.01
60–70	3	0.15	0.015

Age (years)	Frequency	Relative frequency	Frequency density
Total	20	1	—

(b) What is the median age bin?

Solution. The 10th- and 11th-largest observations are both in the 45–50 bin, which is therefore the median bin.

(c) Calculate (an approximation of) the mean age of the politicians.

Solution. Pretending that each person is in the centre of their bin, we have

$$\bar{x} = \frac{1}{20}(1 \times 24 + 3 \times 35 + \cdots + 3 \times 65) = \frac{946.5}{20} = 47.3.$$

B1. For each of the two datasets below, calculate the following summary statistics, or explain why it is not possible to do so: mode; median; mean; number of distinct outcomes; inter-quartile range; and sample variance.

(a) Six packets of Skittles are opened together, and the total number of sweets of each colour is:

Colour	Red	Orange	Yellow	Green	Purple
Number of Skittles	67	71	87	74	62

Solution. The modal colour is Yellow. The number of distinct outcomes is 5.

It's not possible to calculate the median or the quartiles, because, unlike numerical data, the colours can't be put "in order" from smallest to largest.

It's not possible to calculate the mean or sample variance, as these require us to have numerical data that can be "added up", but this can't be done with colours.

(b) Shirt sizes for a university football squad:

Colour	Xtra Small	Small	Medium	Large	Xtra Large
Number of shirts	0	1	6	4	5

*[**Note:** This has been corrected from an earlier version, where the 4 Large and 5 Xtra Large were the wrong way round.]*

Solution. The modal shirt size is medium. The number of distinct outcomes is 4 (we don't quite "Xtra Small", which was not observed in the data).

This time, we can order the data from smallest to largest, even though the data is not numerical. Since $(16+1)/2-8.5$, the median datapoint is the 8th or 9th datapoints, which are Large.

Since $1 + 0.25(16 - 1) = 4.75$ the lower quartile is the 4th or 5th datapoints, which are Medium. Since $1 + 0.75(16 - 1) = 12.25$, the upper quartile is the

12th or 13th datapoints, which are Xtra Large. So we can certainly say that the inner quartiles range from Medium to Xtra Large. We could probably also say that the interquartile range is 3 shirt sizes (Medium, Large, Xtra Large).

Again, because the data is not numerical, we can't add it up, so can't calculate a mean or sample variance.

Group feedback: Make sure your explanation is clear for why we can't calculate a median for the Skittles data but can for the shirts: they key is whether or not the data can be *ordered*.

B2. A summary statistic is informally said to be "robust" if it typically doesn't change much if a small number of outliers are introduced to a large dataset, or "sensitive" if it often changes a lot when a small number of outliers are introduced. Briefly discuss the robustness or sensitivity of the following summary statistics: **(a)** mode; **(b)** median; **(c)** mean; **(d)** number of distinct outcomes; **(e)** inter-quartile range; and **(f)** sample variance.

Solutions.

(a) The mode will typically not change at all if a small number of outliers are introduced, so is robust. (The exception is for data where every observation is likely to be different, so the outliers become "joint modes" along with everything else; but in this case the mode is not a useful statistic in the first place.)

(b) The introduction of outliers will typically only change the median a little bit, by shifting it between different nearby values in the "central mass" of the data. In particular, the size of the outliers won't make any difference at all (only whether they are "high outliers" or "low outliers"). So the median is robust.

(c) The mean can change a lot is outliers are introduced. (Think about the mean net worth of people in you tutorial group, and how it would change if Jeff Bezos or Elon Musk joined your tutorial group.) So the mean is sensitive.

(d) The number of distinct outcomes will only increase by (at most) 1 for each outlier introduced, so is robust.

(e) The interquartile range is robust, for the same reason as the median.

(f) The sample variance is sensitive, for the same reason as the mean.

(You might like to think about situations where it's better to use a robust statistic or better to use a sensitive statistic.)

Group feedback: Remember that "robust" and "sensitive" are general descriptions rather than precise mathematical definitions. So it doesn't matter if you disagree with my opinions provided that you give clear and detailed explanations to back up your opinion.

B3. Let $\mathbf{a} = (a_1, a_2, \dots a_n)$ and $\mathbf{b} = (b_1, b_2, \dots, b_n)$ be two real-valued vectors of the same length. Then the *Cauchy–Schwarz inequality* says that

$$\left(\sum_{i=1}^{n} a_i b_i\right)^2 \leq \left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right).$$

Use the Cauchy–Schwarz inequality to show that the correlation r_{xy} satisfies $-1 \leq r_{xy} \leq 1$.

(*Hint:* Try to prove that $s_{xy}^2 \leq s_x^2 s_y^2$. How does this help?)

Solutions. The first thing we want do is get from the Cauchy–Schwarz inequality

$$\left(\sum_{i=1}^{n} a_i b_i\right)^2 \leq \left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right).$$

to the hint $s_{xy}^2 \leq s_x^2 s_y^2$. We'll do this by making a clever choice for (a_i) and (b_i) in Cauchy–Schwarz that tells us something useful about s_{xy}, s_x^2, and s_y^2.

Recalling the formulas for s_{xy}, s_x^2, and s_y^2,

$$\begin{aligned} s_{xy} &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), \\ s_x^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \\ s_y^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2, \end{aligned}$$

and comparing them with the Cauchy–Schwarz inequality, it looks like taking $a_i = x_i - \bar{x}$ and $b_i = y_i - \bar{y}$ might be useful. Making the substitution, we get

$$\left(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right)^2 \leq \left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right).$$

These are very close to the formulas for s_{xy}, s_x^2, and s_y^2, but are just missing the "$1/(n-1)$"s; what we in fact have is

$$\big((n-1)s_{xy}\big)^2 \leq (n-1)s_x^2 \cdot (n-1)s_y^2.$$

Cancelling $(n-1)^2$ from each side, we have $s_{xy}^2 \leq s_x^2 s_y^2$, and we've proved the hint.

We now want to get from the hint to the desired statement $-1 \leq r_{xy} \leq 1$. Recall the formula for the correlation is

$$r_{xy} = \frac{s_{xy}}{s_x s_y}.$$

We can make the hint look a bit like this dividing both sides by $s_x^2 s_y^2$, to get

$$\frac{s_{xy}^2}{s_x^2 s_y^2} \leq 1.$$

In fact that's the square of the correlation on the left-hand side, so we've shown that $r_{xy}^2 \leq 1$.

Finally, we note that if a number squared is less than or equal to 1, then the number must be between -1 and +1 inclusive. (Numbers bigger than 1 get bigger still when squared; number smalles than -1 become bigger than +1 when squared.) Hence we have shown that $-1 \leq r_{xy} \leq 1$, as required.

Group feedback: There are two halves to this question: first get from the Cauchy–Schwarz inequality to the hint, and second get from the hint to the answer $-1 \leq r_{xy} \leq 1$. Many students struggled with the first half – but you can still try to do the second half. Especially in an exam, look for places where you can get marks for doing part of the question – and it doesn't need to be the *first* part!

B4. A researcher wishes to study the effect of mental health on academic achievement. The researcher will collect data on the mental health of a cohort of students by asking them to fill in a questionnaire, and will measure academic achievement via the students' scores on their university exams. Discuss some of the ethical issues associated with the collection, storage, and analysis of this data, and with the publication of the results of the analysis. Are there ways to mitigate these issues?

(It's not necessary to write an essay for this question – a few short bulletpoints will suffice. There may be an opportunity to discuss these issues in more detail in your tutorial.)

Group feedback: There are no "correct" or "incorrect" answers here, but here are a few things that students in my own tutorials brought up, which may act as a prompt for your own discussions.

- It's important the students/subjects have given their consent for their data to be used this way. It must be "informed consent", where they understand for what purpose the data will be used, how it will be stored, and so on. It must be possible and painless for students to decline to take part.
- Consideration should be given on how to anonymise the data as much as possible – it's not necessary for those analysing the data to know which questionnaire or which exam result belongs to which student, only that the questionnaire and results can be paired up.
- Even if after data is anonymised, care should be taken about whether the students could be worked out from the data. For example, if only one student did a certain combination of modules, their identity could "leak" that way. Perhaps imprecise data, such as classes rather than exact marks, might help while only slightly reducing the usefulness of the data?
- On one hand, it seems like this data should perhaps be deleted once analysis has been carried out, for the privacy of the students. On the other hand, principles of "open science" suggest that the data should be kept – and even publically made available – for other researchers to check the work. There are competing ethical considerations here.
- If correlations are found in the data, care should be taken when publishing the analysis not to wrongly suggest a causation. (Just because X and Y are positively correlated, it doesn't mean that X *causes* Y – or that Y causes X.)

You can probably think of many other things.

Problem Sheet 2

A1. Suppose you toss a coin 10 times. What would you suggest for a sample space **(a)** if you only care about the total number of heads; **(b)** if you care about the result of every coin toss?

Solution.

(a) The number of heads can be any number from 0 to 10, so we should take $\Omega = \{0, 1, 2, \dots, 10\}$. (This sample space contains 11 sample outcomes that are not equally likely.)

(b) The vector of coin outcomes will be something like $(\mathrm{H}, \mathrm{H}, \mathrm{T}, \mathrm{T}, \mathrm{H}, \mathrm{T}, \mathrm{T}, \mathrm{T}, \mathrm{H}, \mathrm{T})$. So our sample space Ω should be the set of all vectors of length 10 whose entries are either H, T; the notation $\Omega = \{\mathrm{H}, \mathrm{T}\}^{10}$ is sometimes used for this. (Note that, by the multiplication principle, this sample space contains $2^{10} = 1024$ sample outcomes that are equally likely.)

A2. Let A, B and C be events in a sample space Ω. Write the following events using only A, B, C and the complement, intersection, and union operations.

(a) C happens but A doesn't.

Solution. This is "C and not A": $C \cap A^{\mathsf{c}}$.

(b) At least one of A, B and C happens.

Solution. This is simply the union $A \cup B \cup C$.

(c) Exactly one of B or C happens.

Solution. One way to write this is to split it up as "'B but not C' or 'C but not B'", which is $(B \cap C^{\mathsf{c}}) \cup (B^{\mathsf{c}} \cap C)$.

An alternative is to split it up as "'B or C' but not 'both B and C'", which is $(B \cup C) \cap (B \cap C)^{\mathsf{c}}$.

You can check these are equal by (for example) using De Morgan's law and the distributive law to expand out the second version.

(d) Exactly two of A, B and C happens.

Solution. I would split this up into "A and B but not C", "A and C but not B", and "B and C but not A" and take the union. This gives

$$(A \cap B \cap C^{\mathsf{c}}) \cup (A \cap B^{\mathsf{c}} \cap C) \cup (A^{\mathsf{c}} \cap B \cap C).$$

There are other equivalent formulations.

A3. Let Ω be a sample space with a probability measure $\mathbb{P}$, and let $A, B \subset \Omega$ be events. State, with brief explanations, whether the following statements are true or false:

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

Solution. False. It is true that *if* $A \subset B$ *then* $\mathbb{P}(A) \leq \mathbb{P}(B)$, but here the implication is the wrong way around.

For a counterexample, consider rolling a dice, and let $A = \{1\}$ and $B = \{2, 3\}$. Then $\mathbb{P}(A) = \frac{1}{6} \leq \frac{2}{6} = \mathbb{P}(B)$, but it's not true that $A \subset B$.

(b) $\mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^{\mathsf{c}}) = \mathbb{P}(A)$.

Solution. True. Note that $(A \cap B) \cup (A \cap B^{\mathsf{c}}) = A$ and that the union is disjoint. (Try drawing a Venn diagram, if this isn't obvious.) The result follows from applying Axiom 3.

(c) $\mathbb{P}(A \cup B) \leq \mathbb{P}(A)$

Solution. False. On the contrary, $A \subset A \cup B$, so the inequality should be the other way round. The same A and B as in part (a) gives a concrete counterexample. The statement would be true with a reversed inequality, or with the union replaced by an intersection.

(d) If A and B are disjoint, then $\mathbb{P}((A \cup B)^{\mathsf{c}}) = 1 - \mathbb{P}(A) - \mathbb{P}(B)$.

Solution. True. From the complement rule, we have $\mathbb{P}((A \cup B)^{\mathsf{c}}) = 1 - \mathbb{P}(A \cup B)$; then from the addition rule for disjoint unions we have $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$. Putting these together gives the result.

A4. What is the value of the following expressions?

(a) $6!$

Solution.

$$6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720.$$

(b) $8^{\underline{4}}$

Solution.

$$8^{\underline{4}} = 8 \times 7 \times 6 \times 5 = 1680$$

(c) $\binom{10}{4}$

Solution.

$$\binom{10}{4} = \frac{10 \times 9 \times 8 \times 7}{4 \times 3 \times 2 \times 1} = 210$$

A5. An urn contains 5 red balls and 7 blue balls. Four balls are drawn from the urn. What is the probability that at least one of the balls is red, if the balls are drawn **(a)** with replacement; **(b)** without replacement?

Solution. This is an "at least one" question, so it will be better to look at the complementary event A^{c} that none of the four balls drawn are red – that is, that they are all blue.

(a) There are $|\Omega| = 12^4 = 20736$ ways to draw four balls with replacement. There are $|A^{\mathsf{c}}| = 2401$ to draw all blue balls. So

$$\mathbb{P}(A) = 1 - \mathbb{P}(A^{\mathsf{c}}) = 1 - \frac{|A^{\mathsf{c}}|}{|\Omega|} = 1 - \frac{2041}{20736} = 0.884.$$

(b) There are $|\Omega| = 12^{\underline{4}} = 11880$ ways to draw four balls without replacement. There are $|A^{\mathsf{c}}| = 7^{\underline{4}} = 840$ to draw all blue balls. So

$$\mathbb{P}(A) = 1 - \mathbb{P}(A^{\mathsf{c}}) = 1 - \frac{|A^{\mathsf{c}}|}{|\Omega|} = 1 - \frac{840}{11880} = 0.929.$$

B1 Starting from just the three probability axioms, prove the following statements:

(a) $\mathbb{P}(\emptyset) = 0$.

Solution. As always, we seek a disjoint union, to allow us to use Axiom 3.

Let A be any event (such as $A = \emptyset$ or $A = \Omega$, for example). Then $A \cup \emptyset = A$, and the union is disjoint – since $\emptyset$ contains no sample points, it certainly can't contain any sample points that are also in A. Then applying Axiom 3, we get $\mathbb{P}(A) + \mathbb{P}(\emptyset) = \mathbb{P}(A)$. Subtracting $\mathbb{P}(A)$ from both sides gives the result.

Alternatively, if you prove part (b) first, you can apply that with $A = \Omega$. Since $\Omega^{\mathsf{c}} = \emptyset$ and Axiom 2 tells us that $\mathbb{P}(\Omega) = 1$, the result follows.

(b) $\mathbb{P}(A^{\mathsf{c}}) = 1 - \mathbb{P}(A)$.

Solution. A very useful and relevant disjoint union is $A \cup A^{\mathsf{c}} = \Omega$. Applying Axiom 3 gives us $\mathbb{P}(A) + \mathbb{P}(A^{\mathsf{c}}) = \mathbb{P}(\Omega)$. But Axiom 2 tells us that $\mathbb{P}(\Omega) = 1$, so $\mathbb{P}(A) + \mathbb{P}(A^{\mathsf{c}}) = 1$. Rearranging gives the result.

B2. Suppose we pick a number at random from the set $\{1, 2, \dots, 2021\}$.

(a) What is the probability that the number is divisible by 5?

Solution. The sample space is $\Omega = \{1, 2, \dots, 2021\}$, and A is the set of numbers up to 2021 that are divisible by 5. Clearly $|\Omega| = 2021$. Further, $|A|$ is the largest integer no bigger than $\frac{2021}{5} = 404.2$, which is 404, as this is how many times 5 "goes into" 2021. Hence

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{404}{2021} = 0.1999,$$

just a tiny bit smaller than $\frac{1}{5}$.

(b) What is the probability the number is divisible by 5 or by 7?

Solution. With the same Ω and A, we now have B being the numbers up to 2021 divisible by 7; so we're looking for $\mathbb{P}(A \cup B)$. As before, $|B|$ is the largest integer no bigger that $\frac{2021}{7} = 288.7$, which is 288. So

$$\mathbb{P}(A \cup B) = \frac{404}{2021} + \frac{288}{2021} - \mathbb{P}(A \cap B).$$

Here, $A \cap B$ is the numbers divisible by both 5 and 7, which is precisely the numbers divisible by $5 \times 7 = 35$. Then $|A \cap B|$ is $\frac{2021}{35} = 57.7$ rounded down. So finally, we have

$$\mathbb{P}(A \cup B) = \frac{404}{2021} + \frac{288}{2021} - \frac{57}{2021} = \frac{635}{2021} = 0.314.$$

B3. In this question, you will have to use the standard two-event form of the addition rule for unions

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

(a) Using the two-event addition rule, show that

$$\mathbb{P}(C \cup D \cup E) = \mathbb{P}(C) + \mathbb{P}(D \cup E) - \mathbb{P}\big(C \cap (D \cup E)\big).$$

Solution. As with the Cauchy–Schwarz question from Problem Sheet 1, the key is to make a good choice for what A and B should be. This time, $A = C$ and $D \cup E$ will work well, since $C \cup (D \cup E) = C \cup D \cup E$. (You can call this "associativity", if you like.) Making that substitution immediately gives us

$$\mathbb{P}(C \cup D \cup E) = \mathbb{P}(C) + \mathbb{P}(D \cup E) - \mathbb{P}\big(C \cap (D \cup E)\big),$$

as required.

(b) Using the two-event addition rule and the distributive law, or otherwise, prove the three-event form of the addition rule for unions:

$$\mathbb{P}(C \cup D \cup E) = \mathbb{P}(C) + \mathbb{P}(D) + \mathbb{P}(E) - \mathbb{P}(C \cap D) - \mathbb{P}(C \cap E) - \mathbb{P}(D \cap E) + \mathbb{P}(C \cap D \cap E).$$

Solution. Let's take the three terms on the right of the equation from part (a) separately.

The first term is $\mathbb{P}(C)$, which is fine as it is.

The second term is $\mathbb{P}(D \cup E)$. This is the probability of the union of two events, so we can use addition rule for the union of two events to get

$$\mathbb{P}(D \cup E) = \mathbb{P}(D) + \mathbb{P}(E) - \mathbb{P}(D \cap E).$$

The third term is $\mathbb{P}\big(C \cap (D \cup E)\big)$. If we use the distributive law, as suggested in the question, we get $C \cap (D \cup E) = (C \cap D) \cup (C \cap E)$, so we want to find $\mathbb{P}\big((C \cap D) \cup (C \cap E)\big)$. But this is another union of two events again, this time with $A = C \cap D$ and $B = C \cap E$. So the addition rule gives

$$\mathbb{P}\big((C \cap D) \cup (C \cap E)\big) = \mathbb{P}(C \cap D) + \mathbb{P}(C \cap E) - \mathbb{P}(C \cap D \cap E),$$

since $(C \cap D) \cap (C \cap E) = C \cap D \cap E$.

Finally, we put this all together, and get

$$\begin{aligned}
&\mathbb{P}(C \cup D \cup E) \\
&\qquad = \mathbb{P}(C) + \big(\mathbb{P}(D) + \mathbb{P}(E) - \mathbb{P}(D \cap E)\big) - \big(\mathbb{P}(C \cap D) + \mathbb{P}(C \cap E) - \mathbb{P}(C \cap D \cap E)\big) \\
&\qquad = \mathbb{P}(C) + \mathbb{P}(D) + \mathbb{P}(E) - \mathbb{P}(C \cap D) - \mathbb{P}(C \cap E) - \mathbb{P}(D \cap E) + \mathbb{P}(C \cap D \cap E).
\end{aligned}$$

which is what we wanted.

B4. Eight friends are about to sit down at random at a round table. Find the probability that

(a) Ashley and Brook sit next to each other, with Chris directly opposite Brook;

Solution. Let Ω be the number of ways the friends can sit around the table. This is an ordering problem, so $\Omega = 8!$.

Let A be the event in the question. What is $|A|$? Well,

- Ashley can sit anywhere, so has 8 choices of seat.
- Brook can sit either directly to Ashley's left or directly to Ashley's right, so has 2 choices of seat.
- Chris must sit directly opposite Brook, so only has 1 choice of seat.
- The remaining five friends can fill up the remaining seats however they like, so have 5, 4, 3, 2, and 1 choices respectively.

Hence $|A| = 8 \times 2 \times 1 \times 5 \times 4 \times 3 \times 2 \times 1$. Thus we get

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{8 \times 2 \times 1 \times 5 \times 4 \times 3 \times 2 \times 1}{8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1} = \frac{2 \times 1}{7 \times 6} = \frac{1}{21}.$$

(b) neither Ashley, Brook nor Chris sit next to each other.

Solution. The sample space Ω is as before. Let's count the outcomes in B, the event in the question.

- Ashley can sit anywhere, so has 8 choices of seat.
- Chris's number of choices will depend on where Brook sits, so we'll have to count their choices together.
  - Brook cannot sit next to Ashley.
  - If Brook sits next-but-one to Ashley, of which there are 2 choices, then Chris has 3 choices: Chris cannot sit on the seat directly between Ashley and Brook, nor directly next to Ashley on the other side, nor directly next to Brook on the other side.
  - If Brook does sits neither next nor next-but-one to Ashley, of which there are 3 choices, then Chris has 2 choices: he cannot sit to the right or left of Ashley, nor to the right or left of Brook.
- The remaining friends have 5, 4, 3, 2, and 1 choices again.

Hence, $|B| = 8 \times (2 \times 3 + 3 \times 2) \times 5 \times 4 \times 3 \times 2 \times 1$. So

$$\mathbb{P}(B) = \frac{|B|}{|\Omega|} = \frac{8 \times (2 \times 3 + 3 \times 2) \times 5 \times 4 \times 3 \times 2 \times 1}{8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1} = \frac{2 \times 3 + 3 \times 2}{7 \times 6} = \frac{12}{42} = \frac{2}{7}.$$

Alternatively, last year a MATH1710 student suggested to me the following rather elegant solution. Suppose the five other friends are already sat at a round table with five chairs. Ashley, then Brook, then Chris will each bring along their own chair, and push into one of the gaps between the friends.

Ashley has 5 gaps to choose from, the Brook will have 6 (Ashley joining the table will have increased the number of gaps by 1), then Chris will have 7, so the total number of ways they can push in is $|\Omega| = 5 \times 6 \times 7$.

To not sit next to each other, Ashley can push in any of the 5 gaps, Brook only has $6 - 2 = 4$ choices (not in the gap directly to the left or right of Ashley), and Chris only has $7 - 4 = 3$ choices (not in the gaps directly to the left or right of Ashley nor the gaps directly to the left or right of Brook – these four gaps are distinct). Hence $|B| = 5 \times 4 \times 3$, and we have

$$\mathbb{P}(B) = \frac{5 \times 4 \times 3}{5 \times 6 \times 7} = \frac{4 \times 3}{6 \times 7} = \frac{12}{42} = \frac{2}{7}.$$

B5. Suppose your tutorial group contains 12 students – you and 11 others. The tutor wishes to choose 4 members of the group to present their work.

(a) How many ways can the tutor choose the presentation group?

Solution. The tutor is sampling 4 items from 12, without replacement (the same person can't be picked twice) and where the order doesn't matter. So this is

$$\binom{12}{4} = 495.$$

(b) How many ways can the tutor choose the presentation group if you are one of the presenters?

Solution. Once the tutor has chosen you, she must pick 3 other students to complete the presentation group out of the other 11 students. So this is

$$\binom{11}{3} = 165.$$

(c) How many ways can the tutor choose the presentation group if you are *not* one of the presenters?

Solution. If the tutor doesn't chose you, she must pick all 4 presenters out of the other 11 students. So this is

$$\binom{11}{4} = 330.$$

(d) Pascal's formula says that

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

Prove Pascal's formula.

Solution. We'd like a "double-counting" argument, where we count the same number in two different ways. Let's follow the breadcrumbs from the first three parts of the question.

Suppose the tutorial group has n students and there will be k presenters. How many ways can the presentation group be chosen?

One way is simply to say that this is $\binom{n}{k}$.

Another way is to separately count the presentation groups that do include you and the presentation groups that don't include you, and add them together. The presentation groups including you require another $k-1$ presenters from the other $n-1$ students, which makes $\binom{n-1}{k-1}$. The presentation groups not including you require all k presenters from the other $n-1$, which makes $\binom{n-1}{k}$. In total, the number of presentation groups is

$$\binom{n-1}{k-1} + \binom{n-1}{k}.$$

Since we've counted the number of presentation groups in two different ways, these expressions must be equal.

Alternatively, you can prove algebraically that

$$\frac{n(n-1)\cdots(n-k+1)}{k(k-1)\cdots 2\cdot 1} = \frac{(n-1)(n-2)\cdots(n-k+1)}{(k-1)(k-2)\cdots 2\cdot 1} + \frac{(n-1)(n-2)\cdots(n-k)}{k(k-1)\cdots 2\cdot 1}$$

(start by making $k!$ the common denominator on the right), but I feel that's not in the spirit of the question.

Problem Sheet 3

A1. Consider dealing two cards (without replacement) from a pack of cards. Which of the following pairs of events are independent?

(a) "The first card is a Heart" and "The first card is Red".

Solution. We have

$$\begin{aligned}\mathbb{P}(\text{first Heart}) &= \frac{13}{52} = \frac{1}{4}\\ \mathbb{P}(\text{first Red}) &= \frac{26}{52} = \frac{1}{2}\\ \mathbb{P}(\text{first Heart and first Red}) &= \mathbb{P}(\text{first Heart}) = \frac{1}{4}.\end{aligned}$$

So $\mathbb{P}(\text{first Heart and first Red}) \neq \mathbb{P}(\text{first Heart})\,\mathbb{P}(\text{first Red})$, and the events are not independent.

(b) "The first card is a Heart" and "The first card is a Spade".

Solution. We have

$$\begin{aligned}\mathbb{P}(\text{first Heart}) &= \frac{13}{52} = \frac{1}{4}\\ \mathbb{P}(\text{first Spade}) &= \frac{13}{52} = \frac{1}{4}\\ \mathbb{P}(\text{first Heart and first Spade}) &= 0.\end{aligned}$$

So $\mathbb{P}(\text{first Heart and first Spade}) \neq \mathbb{P}(\text{first Heart})\,\mathbb{P}(\text{first Spade})$, and the events are not independent.

(c) "The first card is a Heart" and "The first card is an Ace".

Solution. We have

$$\begin{aligned}\mathbb{P}(\text{first Heart}) &= \frac{13}{52} = \frac{1}{4}\\ \mathbb{P}(\text{first Ace}) &= \frac{4}{52} = \frac{1}{13}\\ \mathbb{P}(\text{first Heart and first Ace}) &= \mathbb{P}(\text{first Ace of Hearts}) = \frac{1}{52}.\end{aligned}$$

So $\mathbb{P}(\text{first Heart and first Ace}) = \mathbb{P}(\text{first Heart})\,\mathbb{P}(\text{first Ace})$, and the events are independent.

(d) "The first card is a Heart" and "The second card is a Heart".

Solution. We have

$$\begin{aligned}
\mathbb{P}(\text{first Heart}) &= \frac{13}{52} = \frac{1}{4} \\
\mathbb{P}(\text{second Heart}) &= \frac{13}{52} = \frac{1}{4} \\
\mathbb{P}(\text{first Heart and second Heart}) &= \frac{13 \times 12}{52 \times 51} = \frac{1}{17}
\end{aligned}$$

So $\mathbb{P}(\text{first Heart and second Heart}) \neq \mathbb{P}(\text{first Heart})\,\mathbb{P}(\text{second Heart})$, and the events are not independent.

(e) "The first card is a Heart" and "The second card is an Ace".

Solution. We have

$$\begin{aligned}
\mathbb{P}(\text{first Heart}) &= \frac{13}{52} = \frac{1}{4} \\
\mathbb{P}(\text{second Ace}) &= \frac{4}{52} = \frac{1}{13} \\
\mathbb{P}(\text{first Heart and second Ace}) &= \frac{12 \times 4 + 1 \times 3}{52 \times 51} = \frac{51}{52 \times 51} = \frac{1}{52}
\end{aligned}$$

So $\mathbb{P}(\text{first Heart and second Ace}) = \mathbb{P}(\text{first Heart})\,\mathbb{P}(\text{second Ace})$, and the events are independent.

A2. Three events A, B, C are said to be *pairwise independent* if each pair of events are independent – that is, if A and B are independent, B and C are independent, and A and C are independent. The three events are said to *mutually independent* if they are pairwise independent and also $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\,\mathbb{P}(B)\,\mathbb{P}(C)$.

Consider rolling two dice independently. Let A be the event that the first roll is even, let B be the event that the second roll is even, and let C be the event that the total score is even.

(a) Are A, B, and C pairwise independent?

Solution. First note that the probability that a dice is even is $\frac{3}{6} = \frac{1}{2}$, and the probability the total of two dice is also $\frac{18}{36} = \frac{1}{2}$.

- A and B are independent by assumption.
- Since the sum of two even numbers is even, while even plus odd equals odd, we have that A and C occur if and only if both dice are even, which has probability $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$. Hence $\mathbb{P}(A \cap C) = \mathbb{P}(A) \times \mathbb{P}(C) = \frac{1}{4}$, so A and C are independent.
- For the same reason, V and C occur if and only if both dice are even, which has probability $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$. Hence $\mathbb{P}(B \cap C) = \mathbb{P}(B) \times \mathbb{P}(C) = \frac{1}{4}$, so B and C are independent.

Hence A, B and C are pairwise independent.

(b) Are A, B, and C mutually independent?

Solution. For the same reasons as above, $A \cap B \cap C$ occurs if and only if both dice rolls are even, which happens with probability $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$. Hence, $\mathbb{P}(A \cap B \cap C) = \frac{1}{4}$, but $\mathbb{P}(A)\,\mathbb{P}(B)\,\mathbb{P}(C) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}$. So A, B and C are not mutually independent.

A3. Consider the random variable X with the following PMF:

x	-1	0	0.5	1	2
$p(x)$	0.1	0.3	0.3	0.2	0.1

Find the expectation and variance of X.

Solution. For the expectation,

$$\mathbb{E}X = -1 \times 0.1 + 0 \times 0.3 + 0.5 \times 0.3 + 1 \times 0.2 + 2 \times 0.1 = 0.45.$$

For the variance, we start with

$$\mathbb{E}X^2 = (-1)^2 \times 0.1 + 0^2 \times 0.3 + 0.5^2 \times 0.3 + 1^2 \times 0.2 + 2^2 \times 0.1 = 0.775.$$

Then, using the computational formula,

$$\mathrm{Var}(X) = \mathbb{E}X^2 - \mu^2 = 0.775 - 0.45^2 = 0.5725.$$

*[**Note:** The answers at the bottom of the problem sheet for this were wrong.]*

A4. Consider the random variable X with the following PMF:

x	1	2	4	5	a
$p(x)$	0.1	0.2	0.1	b	c

This random variable has $\mathbb{E}X = 4.3$ and $\mathrm{Var}(X) = 4.61$. Find the values of a, b and c.

Solution. First, a PMF must sum to 1, so

$$1 = 0.1 + 0.2 + 0.1 + b + c,$$

and hence

$$b + c = 0.6. \qquad (1)$$

Second, the expectation is

$$\begin{aligned}\mathbb{E}X &= 1 \times 0.1 + 2 \times 0.2 + 4 \times 0.1 + 5b + ac \\ &= 0.9 + 5b + ac \\ &= 4.3.\end{aligned}$$

Hence

$$5b + ac = 3.4. \qquad (2)$$

Third, the expectation of X^2 is

$$\begin{aligned}\mathbb{E}X &= 1^2 \times 0.1 + 2^2 \times 0.2 + 4^2 \times 0.1 + 5^2 b + a^2 c \\ &= 2.5 + 25b + a^2 c\end{aligned}$$

and this must equal $\mathbb{E}X^2 = \mathrm{Var}(X) + \mu^2 = 4.61 + 3.4^2 = 23.1$. Hence

$$25b + a^2 c = 20.6 \qquad (3).$$

Now we just have to solve the three equations. From (1) we get $b = 0.6 - c$, and substituting into (2) and (3) gives

$$\begin{aligned}(a-5)c &= 0.4 \qquad (4) \\ (a^2-25)c &= 5.6 \qquad (5)\end{aligned}$$

Noting that $a^2 - 25 = (a+5)(a-5)$ is a difference of two squares, we can divide (5) by (4) to get

$$a + 5 = \frac{5.6}{0.4} = 14.$$

So $a = 9$. Substituting this into (4) gives $c = 0.1$, so $b = 0.6 - 0.1 = 0.5$.

A5. The temperature T_C measured in degrees Celsius can be converted to a temperature T_F in degrees Fahrenheit using the formula $T_F = \frac{9}{5}T_C + 32$.

The average daily maximum temperature in Leeds in July is 19.0 °C with a variance of 10.4 °C^2.

(a) What is the average daily maximum temperature in degrees Fahrenheit?

Solution. By linearity of expectation,

$$\mathbb{E}T_F = \mathbb{E}\left(\tfrac{9}{5}T_C + 32\right) = \tfrac{9}{5}\mathbb{E}T_C + 32.$$

So the answer is $\frac{9}{5} \times 19.0 + 32 = 66.2$ °F.

(b) What is the variance of the daily maximum temperature in degrees Fahrenheit squared?

Solution. For the variance,

$$\mathrm{Var}(T_F) = \mathrm{Var}\left(\tfrac{9}{5}T_C + 32\right) = \left(\tfrac{9}{5}\right)^2 \mathrm{Var}(T_C) = \tfrac{81}{25}\,\mathrm{Var}(T_C).$$

So the answer is $\frac{81}{25} \times 10.4 = 33.7$ °F^2.

(c) The average temperature in March is 1.8 °C higher than in February. What is this average difference in degrees Fahrenheit?

Solution. Note that a *change* in temperature in degrees Celsius of δ_C takes us from T_C to $T_C + \delta$. This corresponds to a *change* in temperature in degrees Fahrenheit of

$$\delta_F = \left(\tfrac{9}{5}(T_C + \delta) + 32\right) - \left(\tfrac{9}{5}T_C + 32\right) = \tfrac{9}{5}\delta_C.$$

So, by linearity of expectation again, the answer is $\frac{9}{5} \times 1.8 = 3.2$ °F.

B1. Suppose A and B are a pair of independent events. Show that A and B^{c} are also independent events.

Solution. We need to show that

$$\mathbb{P}(A \cap B^{\mathsf{c}}) = \mathbb{P}(A)\,\mathbb{P}(B^{\mathsf{c}}). \qquad (*)$$

Note that

$$A = (A \cap B) \cup (A \cap B^{\mathsf{c}}),$$

and the union is disjoint, so by Axiom 2,

$$\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^{\mathsf{c}}).$$

Hence, the left-hand side of $(*)$ is

$$\begin{aligned}\mathbb{P}(A \cap B^{\mathsf{c}}) &= \mathbb{P}(A) - \mathbb{P}(A \cap B)\\ &= \mathbb{P}(A) - \mathbb{P}(A)\,\mathbb{P}(B)\\ &= \mathbb{P}(A)(1 - \mathbb{P}(B)),\end{aligned}$$

where, in the second line, we used the fact that A and B are independent.

The right-hand side of $(*)$ is

$$\mathbb{P}(A)\,\mathbb{P}(B^{\mathsf{c}}) = \mathbb{P}(A)(1 - \mathbb{P}(B)),$$

where we've used the complement rule $\mathbb{P}(B^{\mathsf{c}}) = 1 - \mathbb{P}(B)$. Hence, we've shown the left- and right-hand sides of $(*)$ are equal, and we are done.

B2. Chloe cycles to work with probability 0.6 and takes the bus with probability 0.4. She has noticed that she is late 20% of the time when she takes the bus. Her boss notices that, on average, Chloe is late one time per week (that is, once every five days).

(a) What is the probability that Chloe will be late if she cycles to work?

Solution. Let C be the event that Chloe cycles to work and B the event that Chloe takes the bus. Let L be the event that Chloe is late.

- We are told that $\mathbb{P}(C) = 0.6$ and $\mathbb{P}(B) = 0.4$. In particular, B and C make up a partition.
- We are told that $\mathbb{P}(L \mid B) = 0.2$.
- We are told that $\mathbb{P}(L) = \frac{1}{5} = 0.2$.
- We want to find out $\mathbb{P}(L \mid C)$.

The law of total probability tells us that

$$\mathbb{P}(L) = \mathbb{P}(C)\,\mathbb{P}(L \mid C) + \mathbb{P}(B)\,\mathbb{P}(L \mid B).$$

Putting in everything we know, we get

$$0.2 = 0.6\,\mathbb{P}(L \mid C) + 0.4 \times 0.2,$$

which rearranges to $\mathbb{P}(L \mid C) = 0.2$.

(b) Chloe is late today, and her boss suspects this is because she took the bus. What is the probability that this guess is correct?

Solution. This is asking for $\mathbb{P}(B \mid L)$. We know $\mathbb{P}(L \mid B) = 0.2$, and want to switch the order of conditioning, so this is a job for Bayes' theorem! We have

$$\begin{aligned}\mathbb{P}(B \mid L) &= \frac{\mathbb{P}(B)\,(L \mid B)}{\mathbb{P}(L)} \\ &= \frac{0.4 \times 0.2}{0.2} \\ &= 0.4.\end{aligned}$$

*[**Note:** Because $0.2 = \frac{1}{5}$, both types of transport are equally (un)reliable, making this a kind of boring question, which I didn't notice when copying the question from last year's problem sheet. This is a note to remind myself to change the numbers next year.]*

B3. Let Ω be a sample space, let $\mathbb{P}$ be a probability measure on Ω, and fix an event $B \subset \Omega$ with $\mathbb{P}(B) > 0$. Show that the conditional probability $\mathbb{P}(\cdot \mid B)$ is also a probability measure on Ω. That is, show that the conditional probability also satisfies the probability axioms:

1. $\mathbb{P}(A \mid B) \geq 0$ for all events $A \subset \Omega$;
2. $\mathbb{P}(\Omega \mid B) = 1$;
3. For disjoint events $A_1, A_2, \ldots$, we have

$$\mathbb{P}(A_1 \cup A_2 \cup \cdots \mid B) = \mathbb{P}(A_1 \mid B) + \mathbb{P}(A_2 \mid B) + \cdots.$$

Solution. For Axiom 1, we have

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

The numerator is non-negative, by Axiom 1 for the probability measure $\mathbb{P}$, and the denominator is strictly positive, by assumption. Hence, the fraction is non-negative, as required.

For Axiom 2, note that, since $B \subset \Omega$, we have $\Omega \cap B = B$. Hence

$$\mathbb{P}(\Omega \mid B) = \frac{\mathbb{P}(\Omega \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B)}{\mathbb{P}(B)} = 1,$$

as required.

For Axiom 3, note that the dsitributive law tells us that

$$(A1 \cup A_2 \cup \cdots) \cap B = (A_1 \cap B) \cup (A_2 \cap B) \cup \cdots,$$

and the unions remain disjoint. Hence, we have

$$\begin{aligned}
&\mathbb{P}(A_1 \cup A_2 \cup \cdots \mid B) \\
&\quad = \frac{\mathbb{P}\big((A_1 \cup A_2 \cup \cdots) \cap B\big)}{\mathbb{P}(B)} \\
&\quad = \frac{\mathbb{P}\big((A_1 \cap B) \cup (A_2 \cap B) \cup \ldots\big)}{\mathbb{P}(B)} \\
&\quad = \frac{\mathbb{P}(A_1 \cap B) + \mathbb{P}(A_2 \cap B) + \cdots}{\mathbb{P}(B)} \\
&\quad = \frac{\mathbb{P}(A_1 \cap B)}{\mathbb{P}(B)} + \frac{\mathbb{P}(A_2 \cap B)}{\mathbb{P}(B)} + \cdots \\
&\quad = \mathbb{P}(A_1 \mid B) + \mathbb{P}(A_2 \mid B) + \cdots,
\end{aligned}$$

where, in the fourth line we used Axiom 3 for the probability measure $\mathbb{P}(B)$. This proves Axiom 3, and we are done.

B4. Soldiers are asked about their use of illegal drugs, but to protect their privacy and ensure honest answers, a "randomised survey" is used. Each soldier is handed a deck of three cards, picks one of the three cards at random, and responds according to what the card says. The three cards say:

1. "Say 'Yes.' "
2. "Say 'No.' "
3. "Truthfully answer the question 'Have you taken any illegal drugs in the past 12 months?' "

(a) 40% of soldiers respond "Yes". What is the likely proportion of soldiers who have taken illegal drugs in the past 12 months.

Solution. Let C_1, C_2, C_3 be the events that a soldier picks cards 1, 2, or 3 respectively, which have probabilities $\mathbb{P}(C_1) = \mathbb{P}(C_2) = \mathbb{P}(C_3) = \frac{1}{3}$ and make up a partition. Let Y be the event that the soldier answers yes. We know that $\mathbb{P}(Y \mid C_1) = 1$, $\mathbb{P}(Y \mid C_2) = 0$ and $\mathbb{P}(Y \mid C_3) = \mathbb{P}(D)$, where $\mathbb{P}(D)$, which we want to find, is the proportion of soldiers who have taken illegal drugs in the past 12 months. We are also told that $\mathbb{P}(Y) = 0.4$.

The law of total probability tells us that

$$\mathbb{P}(Y) = \mathbb{P}(C_1)\,\mathbb{P}(Y \mid C_1) + \mathbb{P}(C_2)\,\mathbb{P}(Y \mid C_2) + \mathbb{P}(C_3)\,\mathbb{P}(Y \mid C_3).$$

With the information we have, we get

$$0.4 = \tfrac{1}{3} \times 1 + \tfrac{1}{3} \times 0 + \tfrac{1}{3}\,p = \tfrac{1}{3} + \tfrac{1}{3}\,p.$$

Solving this gives $p = \frac{1}{5} = 20\%$.

(b) If a soldier responds "Yes", what is the probability that the soldier has taken illegal drugs in the past 12 months.

Solution. This is asking for $\mathbb{P}(D \mid Y)$. Another one for Bayes theorem:

$$\mathbb{P}(D \mid Y) = \frac{\mathbb{P}(D)\mathbb{P}(Y \mid D)}{\mathbb{P}(Y)}.$$

From the question we know that $\mathbb{P}(Y) = 0.4$. From part (a) we know that $\mathbb{P}(D) = 0.2$. We also know that $\mathbb{P}(Y \mid D) = \frac{2}{3}$, as the soldier will answer Yes is they pick either cards 1 or 3. Hence

$$\mathbb{P}(D \mid Y) = \frac{0.2 \times \frac{2}{3}}{0.4} = \frac{1}{3}.$$

B5. Let X be a random variable. Prove that

(a) $\operatorname{Var}(aX) = a^2 \operatorname{Var}(X)$;

Solution. First, we recall that $\mathbb{E}(aX) = a\mathbb{E}X = a\mu$.

Using the definitional formula, we have

$$\begin{aligned}\operatorname{Var}(aX) = \mathbb{E}(aX - a\mu)^2 &= \mathbb{E}\big(a(X-\mu)\big)^2 \\ &= \mathbb{E}(a^2(X-\mu)^2) = a^2\mathbb{E}(X-\mu)^2 = a^2 \operatorname{Var}(X), \end{aligned} \tag{7.1}$$

where the penultimate equality used the linearity of expectation.

Alternatively: Using the computational formula, we have

$$\begin{aligned}\operatorname{Var}(aX) = \mathbb{E}(aX)^2 - (a\mu)^2 &= \mathbb{E}(a^2X^2) - a^2\mu^2 \\ &= a^2\mathbb{E}X^2 - a^2\mu^2 = a^2(\mathbb{E}X^2 - \mu^2) = a^2 \operatorname{Var}(X). \end{aligned} \tag{7.2}$$

(b) $\operatorname{Var}(X + b) = \operatorname{Var}(X)$.

Solution. First recall that $\mathbb{E}(X + b) = \mathbb{E}X + b = \mu + b$.

Using the definitional formula, we have

$$\operatorname{Var}(X+b) = \mathbb{E}\big((X+b) - (\mu+b)\big)^2 = \mathbb{E}(X-\mu)^2 = \operatorname{Var}(X).$$

(Using the computational formula is possible, but more hassle.)

B6. A gambling game works as follows. You keep tossing a fair coin until you first get a Head. If your first Head comes on the nth coin toss, then you win 2^n pounds.

(a) What is the probability that the first Head is seen on the nth toss of the coin?

Solution. This happens if the first $n-1$ tosses are Tails, with probability $(\frac{1}{2})^{n-1}$, them the nth toss is Heads, with probability $\frac{1}{2}$. Altogether, this is $(\frac{1}{2})^{n-1} \times \frac{1}{2} = (\frac{1}{2})^n$.

(b) Show that the expected winnings from playing this game are infinite.

Solution. The expected winnings are

$$\sum_{n=1}^{\infty} 2^n \times \mathbb{P}(\text{first Head on } n\text{th toss}) = \sum_{n=1}^{\infty} 2^n \times \left(\tfrac{1}{2}\right)^n = \sum_{n=1}^{\infty} 1 = \infty$$

(c) The "St Petersburg paradox" refers to the fact that, despite the fact that the expected winnings from this game are infinite, hardly anybody would be prepared to play this game for, say, £100, and certainly not for £1000. Discuss a few possible "resolutions" to this paradox which could explain why people are unwilling to play this game despite seemingly having infinite expected winnings.

Discussion. One possibility is:

- The people are being irrational, and in fact *should* play the game for £1000.

but I'm not sure anyone *really* thinks that.

Some other possible explanations include:

- The expectation is only infinite if you really could win an extraordinarily large amount of money. Suppose that the person offering the game only has £2^{20}, or just over £1 million. In that case, if the first 20 tosses are all Tails, the opponent gives you all £2^{20} then declares bankruptcy and the game stops. In this more realistic case, your expected winnings are only

$$\sum_{n=1}^{20} 2^n \times \left(\tfrac{1}{2}\right)^n + 2^{20} \times \left(\tfrac{1}{2}\right)^{20} = \sum_{n=1}^{20} 1 + 1 = 21,$$

  or £21; a more reasonable price to pay to play the game.
- The amount of benefit (or "utility") one gets from winning a large amount of money might not be directly proportional to the amount. For example, £200 million might be very nice, but it's not *twice* as nice as £100 million – after all, what else could you really do with the second £100 million. Perhaps the utility of £m scales more logarithmically than linearly, like $\log_2 m$ in some appropriate "happiness units" In that case, the expected *utility* from the game is

$$\sum_{n=1}^{\infty} \log_2(2^n) \times \left(\tfrac{1}{2}\right)^n = \sum_{n=1}^{\infty} n \times \left(\tfrac{1}{2}\right)^n = 2,$$

  happiness units, and you might be willing to pay 2 happiness-units-worth of money to play.
- Normal advice to play games with positive expected winnings only really applies if you can play the game many times (or very similar games). For repeated games, the expected winnings can be interpreted as "the winnings you are likely to get in the long run". For one-off highly unusual games, this doesn't hold, so one needs a different criterion to decide whether to play. (If I was allowed to play this game a million times for £100 a round, but didn't have to settle the money until all one million games had finished, then I would strongly consider playing.)

You can probably come up with other explanations of your own too.