MATH1710 Probability and Statistics I

Matthew Aldridge

University of Leeds, 2022–23

Contents

Schedule

Week 10 (5-9 December):

- **Lecture 19:** The Bayesian idea (Monday 5 December)
- **Lecture 20:** More Bayesian models: (Wednesday 7 December)
- **Problem Sheet 5:** Work through the short and long questions in preparation your tutorial. Deadline for assessed questions: *Monday 12 December*
- **R Worksheet 10:** Law of large numbers to be completed this week
- **R Worksheet 11:** Recap next week's sheet, available now due to the unusually early deadline: Thursday 15 December

Week 9 (28 November – 2 December):

- **Lecture 17:** Exponential distribution and multiple continuous random variables (Monday 28 November)
- **Lecture 18:** Limit theorems: *The lecture on Wednesday 30 November will be cancelled; instead you should learn this material from the online notes and embedded videos*
- **Problem Sheet 5:** Work through the short and long questions in preparation your tutorial next week. Deadline for assessed questions: *Monday 12 December*
- **R Worksheet 9:** Normal distribution to be completed this week; assessed work due Monday 5 December

Week 8 (21–25 November):

- **Tutorials:** *All in-person tutorials cancelled. Online tutorial: Wednesday 23 November at 1000; recording available on Minerva.*
- **Lecture 15:** Continuous random variables (Monday 21 November)
- **Lecture 16:** Normal distribution (Wednesday 23 November)
- **Problem Sheet 4:** Work through the short and long questions in preparation for *the online tutorial.* Deadline for assessed questions: *Tuesday 29 November*
- **R Worksheet 8:** Discrete random variables to be completed this week

Week 7 (14–18 November):

- **Lecture 13:** Multiple random variables (Monday 14 November)
- **Lecture 14:** Expectation and covariance (Wednesday 16 November)
- **Problem Sheet 4:** Work through the short and long questions in preparation for your tutorial next week. Deadline for assessed questions: Monday 28 November.
- **R Worksheet 7:** Discrete distributions to be completed this week; assessed work due Monday 21 November.

Week 6 (7–11 November):

- **Lecture 11:** Binomial and geometric distributions (Monday 7 November)
- **Lecture 12:** Poisson distribution (Wednesday 9 November)
- **Tutorial:** to discuss Problem Sheet 3; check your timetable for details.
- **Problem Sheet 3:** Work through the short and long questions in preparation for your tutorial. Deadline for assessed questions: Monday 14 November.
- **R Worksheet 6:** R Markdown *optional* worksheet for this week.

Week 5 (31 October – 4 November):

- **Lecture 9:** Discrete random variables (Monday 31 October)
- **Lecture 10:** Expectation and variance (Wednesday 2 November)
- **Problem Sheet 3:** Work through the short and long questions in preparation for your tutorial in Week 6. Deadline for assessed questions: Monday 14 November.
- **R Worksheet 5:** Plots II to be completed this week. Deadline for assessed questions: Monday 7 November.

Week 4 (24–28 October):

- **Lecture 7:** Independence and conditional probability (Monday 24 October)
- **Lecture 8:** Two theorems on conditional probability (Wednesday 26 October)
- **Tutorial:** to discuss Problem Sheet 2; check your timetable for details.
- **Problem Sheet 2:** Work through the short and long questions in preparation for your tutorial. Deadline for assessed questions: Monday 31 October.
- **R Worksheet 4:** Plots I to be completed this week.

Week 3 (17–21 October):

- **Lecture 5:** Classical probability I (Monday 17 October)
- **Lecture 6:** Classical probability II (Wednesday 19 October)
- **Problem Sheet 2:** Work through the short and long questions in preparation for your tutorial in Week 4. Deadline for assessed questions: Monday 31 October.

- **R Worksheet 3:** Data in R to be completed this week. Deadline for assessed questions: Monday 24 October.

Week 2 (10–14 October):

- **Lecture 3:** Sample spaces and events (Monday 10 October)
- **Lecture 4:** Probability (Wednesday 12 October)
- **Tutorial:** to discuss Problem Sheet 1; check your timetable for details.
- **Problem Sheet 1:** Work through the short and long questions in preparation for your tutorial. Deadline for assessed questions: Monday 17 October.
- **R Worksheet 2:** Vectors to be completed this week.

Week 1 (3–7 October):

- **Lecture 1:** Summary statistics (Monday 3 October)
- **Lecture 2:** Data visualisation (Wednesday 5 October)
- **Problem Sheet 1:** Work through the short and long questions in preparation for your tutorial in Week 2. Deadline for assessed questions: Monday 17 October.
- **R Worksheet 1:** R basics to be completed this week.

About MATH1710

Organisation of MATH1710

This module is **MATH1710 Probability and Statistics I**. (It is possible to take this module as half of **MATH2700 Probability and Statistics for Scientists**, but I am not aware that any students are enrolled on MATH2700 this year – please let me know if you are.)

This module lasts for 11 weeks from 3 October to 16 December 2022. The exam will take place between 16 and 27 January 2023.

The module leader, the lecturer, and the main author of these notes is Dr Matthew Aldridge (you can call me "Matt" or "Dr Aldridge", pronounced "*old*-ridge").

Lectures

The main way you will learn new material for this module is by attending lectures. There are two lectures per week. Because this is a very large class, you are split into two groups for lectures:

- Group 1: Mondays at 1200 and Wednesdays at 1600
- Group 2: Mondays at 1500 and Wednesdays at 1500

All lectures are in Roger Stevens LT 20. Check your timetable to see which group you are in.

I recommend taking your own notes during the lecture. This website will keep brief notes from the lectures, summarising the main definitions and theorems, but will not reflect all the details I say and write during the lectures. Lectures will go through material quite quickly and the material may be quite difficult, so it's likely you'll want to spend time reading through your notes after the lecture.

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

I am very keen to hear about errors in the notes, mathematical, typographical or otherwise. Please email me if think you may have found any.

Attendance at lectures is compulsory.

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%. Deadlines are 2pm on Mondays, although I'd personally recommend completing and submitting the work in the previous week.

Problem Sheet	Lectures covered	Deadline for assessed work
1	1 and 2	Monday 17 October (Week 3)
2	3–6	Monday 31 October (Week 5)
3	7–10	Monday 14 November (Week 7)
4	11–14	Monday 28 November (Week 9)
5	15–18	Monday 12 December (Week 11)

An informal Problem Sheet 6 covering material from Lectures 19 and 20 will be available; Lectures 21 and 22 are revision lectures with no new material.

Assessed questions should be submitted in PDF format through Gradescope. (Further Gradescope details will follow.) Most students choose to hand-write their solutions on paper 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.

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; you should check your timetable to see when and where your tutorial group meets.

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. Because of the much smaller groups, the tutorials are the most valuable type of teaching on the module; you should make sure you attend, and you should be well prepared to ensure you make the most of the opportunity.

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 4 hours per problem sheet.* This is a long time, but you shouldn't expect to be able to answer the hardest questions on a problem sheet with making multiple attempts. You don't have to wait until all lectures in a section are complete until starting to work on some of the questions – this is particularly important for students with Monday tutorials. 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.

Attendance at tutorials is compulsory.

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 academia 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 few questions for assessment, which will be due by 2pm Monday the following week (except the last one). Each of these is worth 3% of your mark for a total of 15%. You will submit your answers through a Microsoft Form (details to follow later). I recommend spending one hour per week on the week's R worksheet, plus one extra hour if there are assessed questions that week.

Week	Worksheet	Deadline for assessed work
1	R basics	—
2	Vectors	—
3	Data in R	Monday 24 October (Week 4)
4	Plots I: Making plots	—
5	Plots II: Making plots better	Monday 7 November (Week 6)
6	RMarkdown (optional)	—
7	Discrete distributions	Monday 21 November (Week 8)
8	Discrete random variables	—

Week	Worksheet	Deadline for assessed work
9	Normal distribution	Monday 5 December (Week 10)
10	Law of large numbers	—
11	Recap	Thursday 15 December (Week 11)

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".

Attendance at R troubleshooting drop-in sessions is optional.

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 leader, 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 I'm happy to discuss any statistics-related issues or questions you have.

I currently plan two optional "office hours" drop-in session per week:

- Thursdays from 1400 to 1500 in Roger Stevens LT 7
- Thursdays from 1600 to 1700 in Roger Stevens LT 17

Although only the second of these appears on your timetable, you are equally welcome at either. Depending on attendance levels, I may change arrangements as term continues. If neither time is possible, you may email me to book a time to talk to me.

Attendance at "office hours" drop-in sessions is optional. You should prioritise mandatory sessions (like lectures or tutorials, such as for LUBS1940 Economics for Management) over this optional session.

Time management

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

- **Lectures:** 2 hours per week, plus 1 hour per week reading through notes.
- **Problem sheets:** 4 hours per problem sheet, plus 1 extra hour for writing up and submitting answers to assessed questions.

- **R worksheets:** 1 hour per week, plus 1 extra hour if there are assessed questions.
- **Tutorials:** 1 hour every other week.
- **Revision:** 15 hours total at the end of the module.
- **Exam:** 2 hours.

That 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...?

There are over 420 students on this module. If each student emails me once a week, and if each email takes me 10 minutes to read and respond, that will take more than 14 hours of my time every day. Generally, it's much better to come to speak to me at the "office hours" drop-in session or, if it will be very quick, before or after a lecture.

- *I don't understand something in the notes or on a problem sheet*: Come to office hours, or ask your tutor in your next tutorial.
- *I'm having difficulties with R:* In Weeks 2 or 3, you should attend an 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 talk to me before/after lectures.
- *I have an admin question about arrangements for my tutorial:* Contact your tutor.
- *I have an admin question about general arrangements for my programme as a whole:* Contact the Student Information Service 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:* You can email me about this.
- *Due to truly exceptional and unforeseeable personal circumstances I require an extension on or exemption from assessed work:* You can apply by filling in the mitigating circumstances form at this link. Neither I nor your tutor can unilaterally offer an extension or exemption, so please don't ask. (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”. I’ll assume you have some basic school-level maths knowledge, but I won’t assume you’ve studied probability or statistics in detail before (although I 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.

- **Exploratory data analysis** [2 lectures]: Summary statistics, data visualisation
- **Probability** [16 lectures]:
  - Probability with events: Probability spaces, probability axioms, examples and properties of probability, “classical probability” of equally likely events, independence, conditional probability, Bayes’ theorem [6 lectures]
  - 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 [10 lectures]
- **Bayesian statistics** [2 lectures]: Bayesian framework, Beta prior, normal–normal model
- Summary and revision [2 lectures]

You’ll notice that this module is heavier on the “Probability” than the “Statistics” of its title. MATH1712 Probability and Statistics II, on the other hand, which many students on this module will take next semester, is almost entirely “Statistics”.

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. These books are also a good place to look if you want extra exercises and problems for revision.

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, we will only taste a brief introduction, but if you want a book, I recommend:

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

For R, there are many excellent resources online.

(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, and were edited and updated in 2022. They are based in part on previous notes by Dr Robert G Aykroyd and Prof Wally Gilks. Dr Jason Susanna 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. If you have accessibility difficulties with these notes, contact me.

Part I: Exploratory data analysis

Chapter 1

Summary statistics

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?
- *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:

```r
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 the answer here only has one part anyway). Plots produced by R are displayed in these notes 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 Statistics of centrality

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:

- **Statistics of centrality**, which tell us where the "middle" of the data is.
- **Statistics 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. 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 *nine* different settings for choosing quantiles! – we will always just use R's default choice.)

- The **median** is the $\frac{1}{2}$-quantile $q(\frac{1}{2})$, which is $q(\frac{1}{2}) = x_7 = 60$ for this data.
- The **minimum** is the 0-quantile $q(0)$, which is $q(0) = x_1 = 59$ for this data.
- The **maximum** is the 1-quantile $q(1)$, which is $q(1) = x_{13} = 63$ for this data
- The **lower quartile** (that's "quartile", as in "quarter" – not "quantile") is the $\frac{1}{4}$-quantile $q(\frac{1}{4})$, which is $q(\frac{1}{4}) = x_4 = 59$ for this data.
- The **upper quartile** is the $\frac{3}{4}$-quantile $q(\frac{3}{4})$, which is $q(\frac{3}{4}) = x_{10} = 62$ for this data.

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 lower quartile 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
```

```
quantile(jul$temp, probs = 1 / 4)
```

```
## 25%
##  15
```

1.4 Statistics of spread

Some measures of spread are:

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

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

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 usually takes fewer presses of calculator buttons to compute the sample variance with this formula rather than the definitional formula. (But make sure you keep enough decimal points in $\bar{x}^2$.)

Going back to our weather data in R, we can find the sample variance of the October weather or the interquartile range of the July weather.

```
var(oct$temp)
```

```
## [1] 2.862366
```

```
IQR(jul$temp)
```

```
## [1] 3
```

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.

Chapter 2

Data visualisations

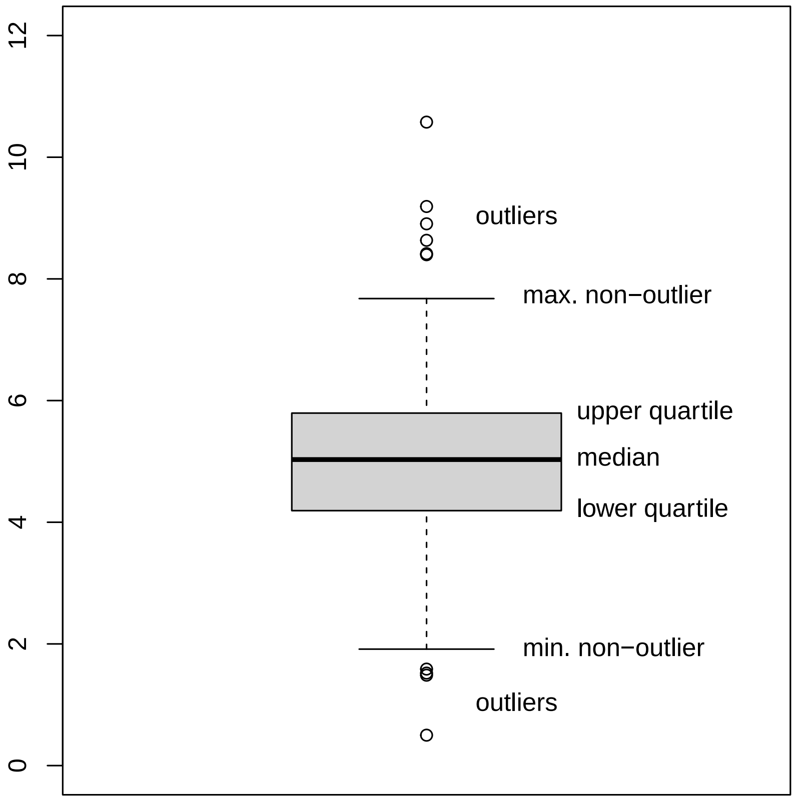
Data visualisations – drawings or graphs based on data – can help us to understand the "shape" of a dataset as part of exploratory data anlaysis. In this lecture, we'll look at three types of data visualisation.

2.1 Boxplots

A **boxplot** is a useful way to illustrate numerical 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.

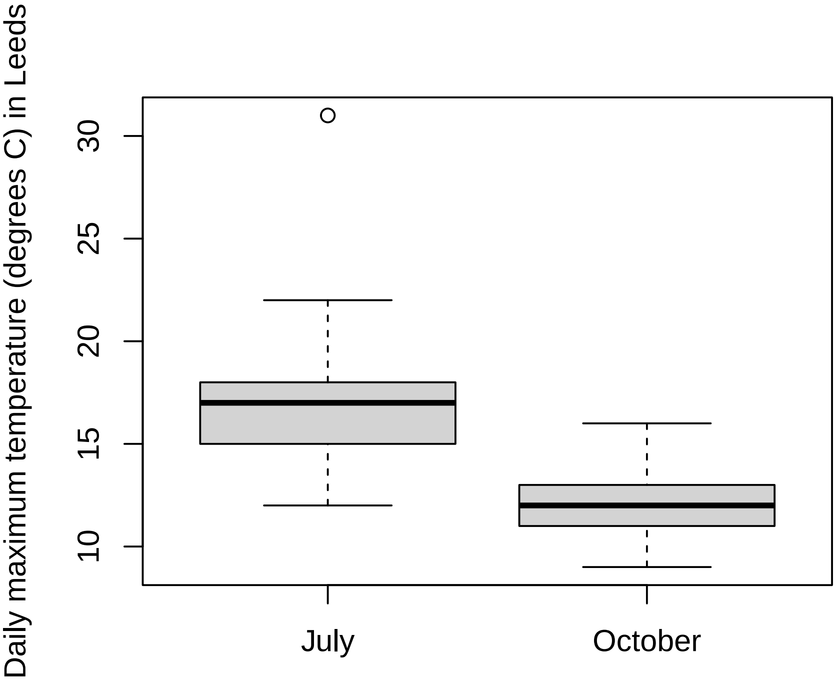
A boxplot is drawn as follows:

- 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 $q(\frac{1}{4}) - 1.5 \times \mathrm{IQR}$ or greater than $q(\frac{3}{4}) + 1.5 \times \mathrm{IQR}$ is an outlier.)
- Out from the two previous boxes, draw "whiskers" to the minimum and maximum non-outlier datapoints.



When we have multiple datasets, drawing boxplots next to each other can help us to compare the datasets. Here are two boxplots from the July and October temperature data we used in the last lecture. 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.)

2.2 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.

Which bin would you say is the most popular – that is, the "modal" bin? 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	

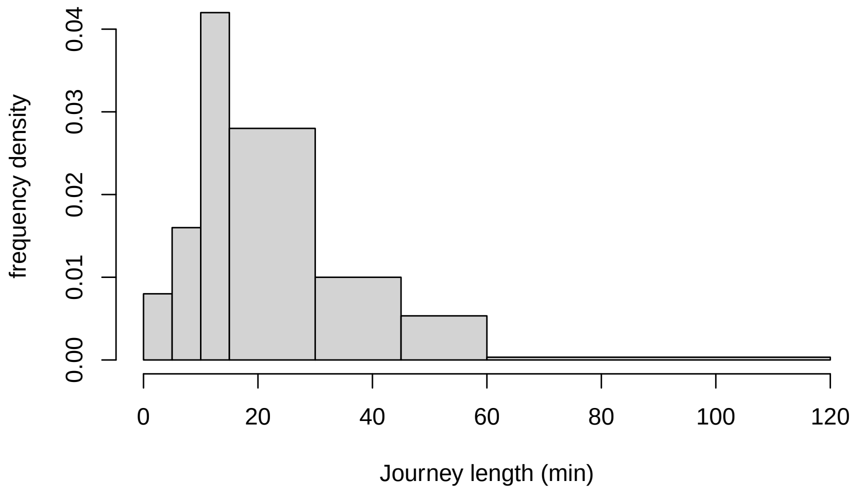
In the first row, for example, the relative frequency is 0.04 and the size of the bin is 5 minutes, so the frequency density is $0.04/5 = 0.008$. We now see that the modal bin – the bin with the highest frequency *density* – is in fact the "10–15 minutes" bin. This bin has somewhat fewer datapoints that the "15–30 minutes" bin, but they're squashed into a much smaller bin.

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, where bars are 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 in the equal-width-bins case.

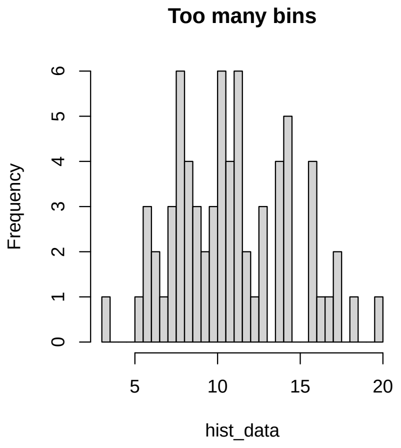
Here is a histogram for our journey-time data:

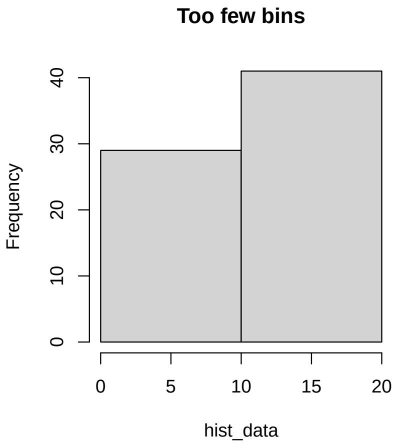
```r
journeys <- read.csv("https://mpaldridge.github.io/math1710/data/journeys.csv")
bins <- c(0, 5, 10, 15, 30, 45, 60, 120)

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



Often we draw histograms because the data was collected in bins in the first place. 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 wide bins will mean we lose detail (see right).

```
set.seed(2172)
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 = 2,  main = "Too few bins")
```





We can also calculate some summary statistics even when we have binned data. We mentioned the mode earlier, where the modal bin is the bin of highest frequency density.

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.

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 could 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

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

2.3 Scatterplots

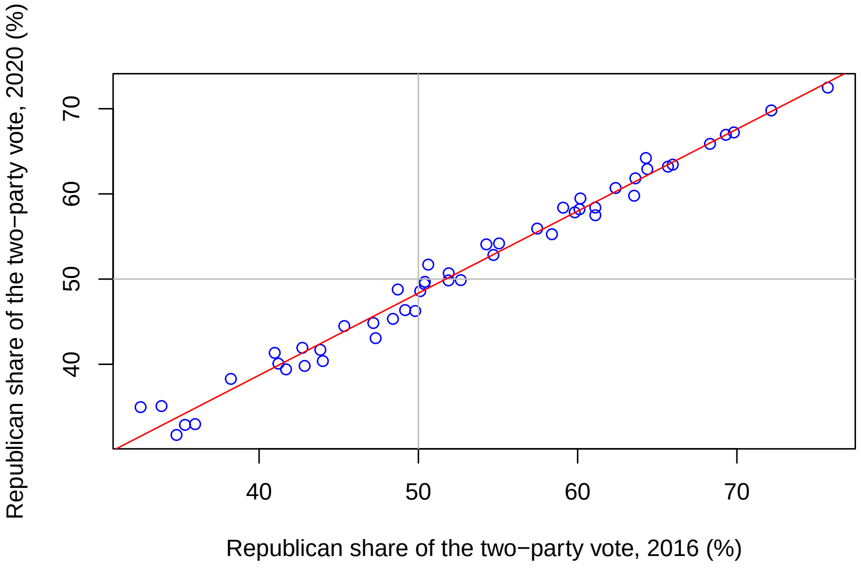
Often, more than one piece of data is collected from each subject, and we wish to compare that data, to see if there is a relationship between the variables.

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 $s_x = \sqrt{s_x^2}$ and $s_y = \sqrt{s_y^2}$ 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

- Boxplots show the shape of numerical data, and can compare different datasets.
- Histograms show the shape of binned data.
- Scatterplots show the relationship between two datasets.

Problem Sheet 1

You can download this problem sheet as a PDF file

This is Problem Sheet 1, which covers material from Lectures 1 and 2 of the notes. You should work through all the questions on this problem sheet in preparation for your tutorial in Week 2. Questions C1 and C2 are assessed questions, and are due in by **2pm on Monday 17 October**. I recommend spending about 4 hours on this problem sheet, plus 1 extra hour to neatly write up and submit your answers to the assessed questions.

A: Short questions

The first three questions are **short questions**, which are intended to be mostly not too difficult. Short questions usually follow directly from the material in the lectures. Here, you 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" bin, for example, means from one's 18th birthday to the moment before one's 30th birthday, so lasts 12 years.)

Age (years)	Frequency	Relative frequency	Frequency density
18–30	1		

Age (years)	Frequency	Relative frequency	Frequency density
30–40	3		
40–45	4		
45–50	5		
50–55	3		
55–60	1		
60–70	3		
Total	20	1	—

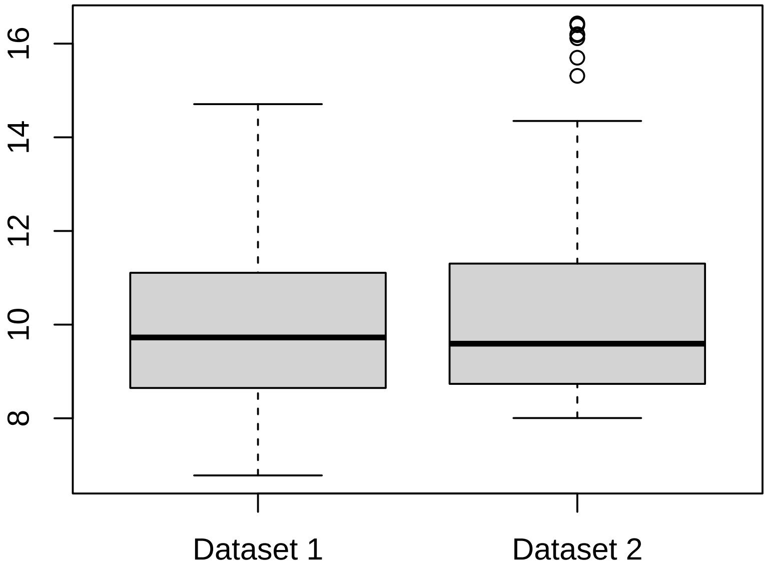
(a) Complete the table by filling in the relative frequency and frequency densities.

(b) What is the median age bin?

(c) What is the modal age bin?

(d) Calculate (the standard approximation of) the mean age of the politicians.

A3. Consider the two datasets illustrated by the boxplots below. Write down some differences between the two datasets.



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. You may not be able to solve all of these questions, although you should make multiple attempts to do so. 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

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

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

(a) By making a clever choice of (a_i) and (b_i) in the Cauchy–Schwarz inequality, show that $s_{xy}^2 \leq s_x^2 s_y^2$.

(b) Hence, show that the correlation r_{xy} satisfies $-1 \leq r_{xy} \leq 1$.

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 small hint in your tutorial.

The deadline for submitting your solutions is **2pm on Monday 17 October** at the beginning of Week 3. Submission will be via Gradescope, which you can access via Minerva. 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. Your work will be marked by your tutor and returned on Monday 24 October, 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:

$$\begin{gathered}1.306,\ 1.301,\ 1.290,\ 1.266,\ 1.268,\ 1.302,\\ 1.317,\ 1.304,\ 1.284,\ 1.268,\ 1.247,\ 1.215.\end{gathered}$$

(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 sort of 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 $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ (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) 2.40, (c) 2.40, (d) —. **A2.** (a) —, (b) 45–50, (c) 45–50, (c) 47.3. **A3.** —

Part II: Probability

Chapter 3

Sample spaces and events

3.1 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.

3.2 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 3.1. 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 3.2. 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 3.3. 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 3.4. 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.

3.3 Set theory

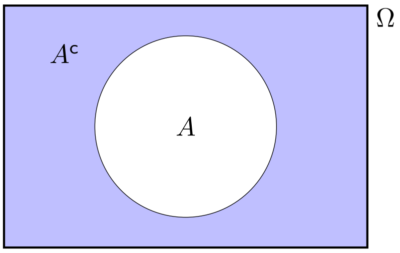
Since we've now defined events as being sets – specifically, subsets of the sample space Ω – it will be useful to mention a little set basic theory here.

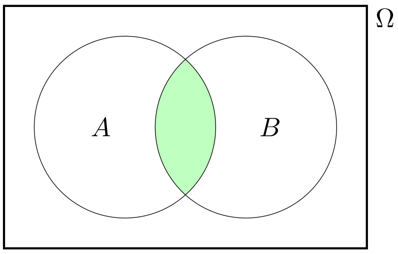
First, there are ways we can build new sets (or events) out of old. It's fine to just read the words and look at the pictures for these definitions, but those who want to read the equations too will need to know this:

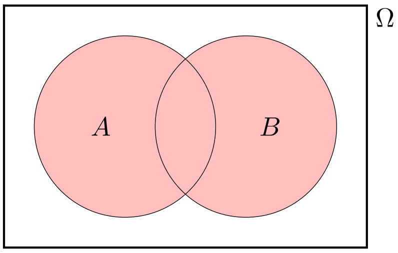
- $\omega \in A$ means "ω is in A" or "ω is an element of A", while $\omega \notin A$ means the opposite, that ω is *not* in A;
- a colon $:$ in the middle of set notation should be read as "such that";
- so $\{\omega \in \Omega : \text{fact about } \omega\}$ should be read as "the set of sample points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

Definition 3.1. 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 3.5. 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.

You might think that if two events are disjoint, then it would be reasonable to find the probability of their union – that is, the probability that one (and, by necessity, only one) of them happens – you can just add the two separate probabilities together. This will be another of our "axioms" of probability.

There are a few rules about ways you can combine the complement, intersection and union operations. These are ways of building new events from old.

- 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 "multiply out of the brackets" with sets:

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

The first says that if you are eating a burger with fries or salad, then you're eating a burger with fries or eating a burger with salad. The second is a bit less intuitive, I find, but it's clear that if A is true then the first of each of the terms on the right is true, while if both B and C are true then the second of each of the terms on the right is true.

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

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

  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. (Augustus De Morgan was a British mathematician of the 19th century who did important work in logic.)

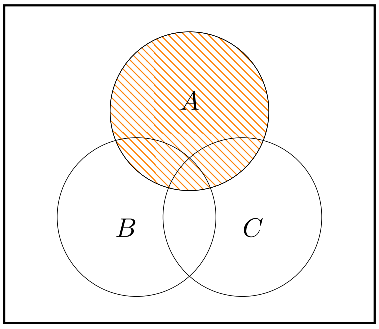
For this module, these mostly count as "common sense" – but if you ever do need to prove one of these statements (or a similar one), one way is to use a Venn diagram.

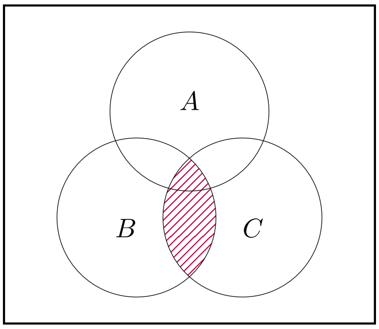
Let's prove the second distributive law,

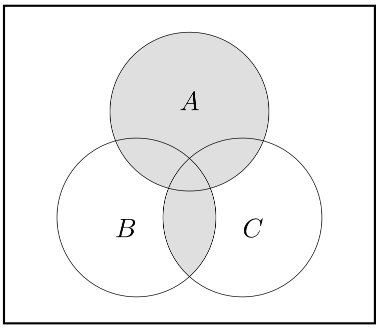
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C),$$

with a Venn diagram as an example.

We can build the left-hand side of the law as:

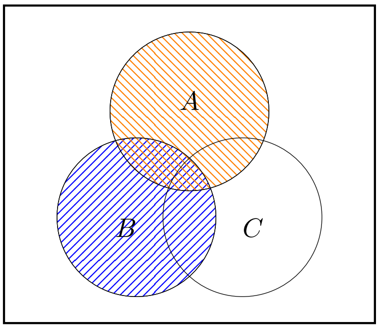


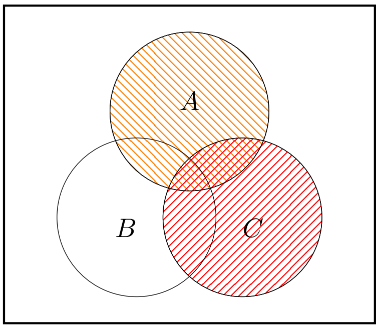


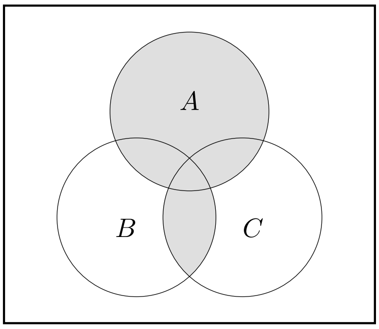


The left-hand figure is A, the middle figure is $B \cap C$, and the right-hand figure is union of these, $A \cup (B \cap C)$.

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

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" $A \cup B$.

Chapter 4

Probability

4.1 Probability axioms

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

Definition 4.1. 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 $A_1, A_2, ...$ 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 very precise should note that an infinite sequence in Axiom 3 must be "countable"; that is, indexed by the natural numbers $1, 2, 3. ....$)

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 4.1. 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

$$\begin{aligned} \mathbb{P}(\emptyset) &= 0 & \mathbb{P}(\{\mathrm{H}\}) &= p \\ \mathbb{P}(\{\mathrm{T}\}) &= 1 - p & \mathbb{P}(\{\mathrm{H}, \mathrm{T}\}) &= 1. \end{aligned}$$

Let's check that the axioms hold:

1. Since $0 \leq p \leq 1$, all the probabilities 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 $\{\mathrm{H}\} \cup \{\mathrm{T}\} = \{\mathrm{H}, \mathrm{T}\}$, where we see that

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

   as required.

Example 4.2. 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 lecture, and prove that the classical probability measure does indeed satisfy the axioms

4.2 Properties of probability

The axioms of Definition 4.1 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 merely "seem likely to be true" or "are common sense" – we can only use the three axioms together with strict logical deductions and nothing else.

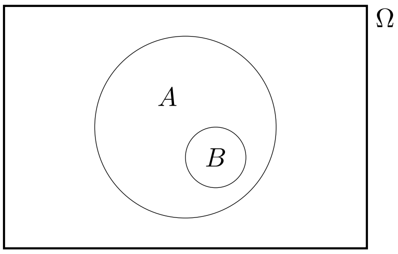
Theorem 4.1. *Let Ω be a sample space with a probability measure $\mathbb{P}$. Then we have the following:*

1. $\mathbb{P}(\emptyset) = 0$.
2. $\mathbb{P}(A^\mathsf{c}) = 1 - \mathbb{P}(A)$ *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 second 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.

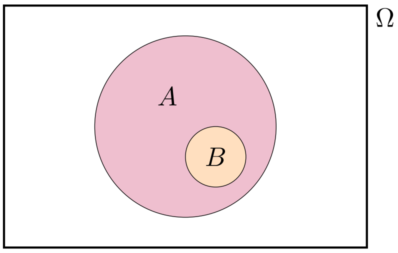
Proof. The key with most of these "prove from the axioms" problems is to think of a way to write the relevant events as part of a *disjoint* union, then use Axiom 3. Statements 1 and 2 are exercises for you on Problem Sheet 2. We'll start with the third statement.

Here, since B is a subset of A, meaning that B is entirely inside 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

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



Applying Axiom 3 to this disjoint union gives

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

We're happy to see the term on the left-hand side and the first 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 the probability of any event is greater than or equal to 0, so in particular. $\mathbb{P}(A \cap B^\mathsf{c}) \geq 0$. Hence

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

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 we have $A \subset \Omega$, so the third statement tells us that $\mathbb{P}(A) \leq \mathbb{P}(\Omega)$. But Axiom 2 tells us that $\mathbb{P}(\Omega) = 1$, so $\mathbb{P}(A) \leq 1$ and we are done. □

4.3 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 4.2. *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.

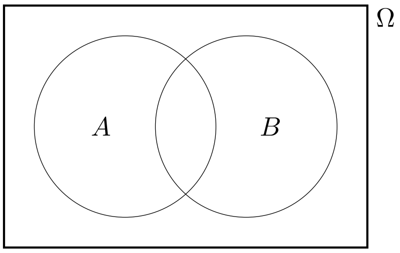
Theorem 4.3. *Let $A, B \subset \Omega$ be two events. Then*

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

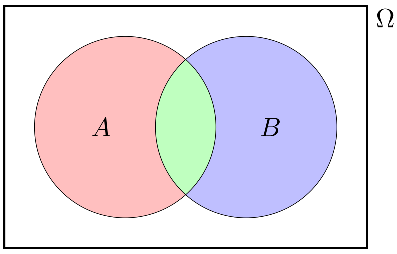
You may have seen this result before. You've perhaps justified it by saying something like this: "We can add the two probabilities together, except now we've double-counted the overlap, so we have to take the probability of that away." Maybe you drew a Venn diagram. That's OK as a way to remember the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

(The following proof has been updated to match how I taught it in Wednesday's lecture.)

Proof. The problem here is that A and B are not (in general) disjoint, so we can't apply Axiom 3.



Instead, let's split this up into the three disjoint bits: "A but not B" $A \cap B^{\mathsf{c}}$, "B but not A" $B \cap A^{\mathsf{c}}$, and "both" $A \cap B$.



Now we can write A, B and $A \cup B$ in terms of these disjoint bits.

$$\begin{aligned} A &= (A \cap B^{\mathsf{c}}) \cup (A \cap B) && (4.1)\\ B &= (B \cap A^{\mathsf{c}}) \cup (A \cap B) && (4.2)\\ A \cup B &= (A \cap B^{\mathsf{c}}) \cup (B \cap A^{\mathsf{c}}) \cup (A \cap B), && (4.3) \end{aligned}$$

with all the unions on the right-hand side being disjoint. Applying Axiom 3 to them all gives

$$\begin{aligned} \mathbb{P}(A) &= \mathbb{P}(A \cap B^{\mathsf{c}}) + \mathbb{P}(A \cap B) && (4.4)\\ \mathbb{P}(B) &= \mathbb{P}(B \cap A^{\mathsf{c}}) + \mathbb{P}(A \cap B) && (4.5)\\ \mathbb{P}(A \cup B) &= \mathbb{P}(A \cap B^{\mathsf{c}}) + \mathbb{P}(B \cap A^{\mathsf{c}}) + \mathbb{P}(A \cap B). && (4.6) \end{aligned}$$

Here, (4.6) is looking good, but we need to get rid of the awkward $\mathbb{P}(A \cap B^{\mathsf{c}})$ and $\mathbb{P}(B \cap A^{\mathsf{c}})$ terms. We can do that be rearranging (4.4) and (4.5) to get

$$\begin{aligned} \mathbb{P}(A \cap B^{\mathsf{c}}) &= \mathbb{P}(A) - \mathbb{P}(A \cap B) && (4.7)\\ \mathbb{P}(B \cap A^{\mathsf{c}}) &= \mathbb{P}(B) - \mathbb{P}(A \cap B). && (4.8) \end{aligned}$$

Substituting these into (4.6) gives

$$\begin{aligned} \mathbb{P}(A \cup B) &= \mathbb{P}(A) - \mathbb{P}(A \cap B) + \mathbb{P}(B) - \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B) && (4.9)\\ &= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B), && (4.10) \end{aligned}$$

as required. □

Example 4.3. *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

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

Summary

- The axioms of probability are (1) $\mathbb{P}(A) \geq 0$; (2) $\mathbb{P}(\Omega) = 1$; and (3) that for disjoint events $A_1, A_2, ...$, 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 5

Classical probability I

5.1 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 5.1. 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 lecture and the next, we’ll look at some different ways in which we can count the number of outcomes in common events and sample spaces.

Example 5.1. *We roll a dice. What is the probability we get at least 5?*

The sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$, with $|\Omega| = 6$. The event that we roll at least 5 is $A = \{5, 6\}$, with $|A| = 2$. Hence

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{2}{6} = \frac{1}{3}.$$

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

Theorem 5.1. *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 it satisfies the three axioms in Definition 4.1.

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.

□

5.2 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 × 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
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 5.2. *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 $A = \{\mathrm{HHHHH}, \mathrm{TTTTT}\}$, 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 5.3. *Four dice are rolled. What is the probability we get at least one 6?*

Here, Ω is the set of all possible sequences of four dice rolls. Clearly $|\Omega| = 6^4 = 1296$.

The event 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 $\mathbb{P}(A) = 1 - \mathbb{P}(A^{\mathsf{c}})$, but in "at least one" questions, it's often easier to count $|A^{\mathsf{c}}|$ than to count $|A|$.

Here, since A is the set of all dice roll sequences with at least one 6, then A^{c} is the set of dice roll sequence without any 6s at all. This means all four dice must have rolled a 1, 2, 3, 4 or 5. Since each of the four dice rolls has five possibilities, this means that $|A^{\mathsf{c}}| = 5^4 = 625$.

Putting this together, we see that

$$\mathbb{P}(A) = 1 - \mathbb{P}(A^{\mathsf{c}}) = 1 - \frac{|A^{\mathsf{c}}|}{|\Omega|} = 1 - \frac{625}{1296} = \frac{671}{1296} \approx 0.518.$$

So there's about a 52% chance we get at least one 6.

5.3 Sampling with and without replacement

Example 5.4. *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 ball 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**. This is still k numbers multiplied together, but decreasing by 1 each time down from n. The final number in the product is the number of choices in the kth an final round: this is the original n items minus the $k-1$ items sampled in the previous $k-1$ rounds; so the final number is $n-(k-1) = n-k+1$, not $n-k$. A notation point: Notice that the subscript is underlined in the falling factorial; other notation sometimes used includes $(n)_k$, $P(n,k)$, or ${}^nP_k$.

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

Chapter 6

Classical probability II

We continue looking at the classical probability $\mathbb{P}(A) = |A|/|\Omega|$, by looking at ways to enumerate Ω and A. Last time we saw:

- The multiplication principle: n_1 choices followed by n_2 choices, …, up to n_k choices gives $n_1 \times n_2 \times \cdots \times n_k$ 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.

6.1 Ordering

Example 6.1. *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 the lecturer is not going to mark the same paper twice. There are $n-2$ choices for the third exam paper. And so on, until she has 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.

The factorial $n!$ gets very large very quickly. **Stirling's formula** gives the approximation $n! \approx \sqrt{2\pi n}\,\mathrm{e}^{-n}\,n^n$.

Example 6.2. Suppose you shuffle a pack of cards. The resulting ordering of the deck has 52! possibilities. This is an unimaginably huge number – the exact value to 3 significant figures is

$$52! = 8.07 \times 10^{67},$$

while Stirling's formula gives the approximation

$$52! \approx \sqrt{2\pi \times 52} \times \mathrm{e}^{-52} \times 52^{52} = 8.05 \times 10^{67}.$$

This is an 8 followed by 67 zeroes.

If every person on the planet (very roughly 10^{10}) had shuffled a deck of cards one million (10^6) times a second for the entire lifetime of the universe (roughly 10^{17} seconds), they could only expect to have got through about 10^{33} shuffles. 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.

If we take the ratio of a bigger factorial $n!$ over a smaller factorial $j!$, we get lot of cancellation,

$$\begin{aligned}\frac{n!}{j!} &= \frac{n(n-1)\cdots(j+1)j(j-1)\cdots 1}{j(j-1)\cdots 1} \\ &= n(n-1)\cdots(j+1),\end{aligned}$$

because the last part of the product in the numerator cancels with the whole of the denominator. Replacing j with $n-k$, this gives

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

This gives a way of writing the falling factorial as the ratio of two (normal) factorials, which can sometimes be useful.

6.2 Sampling without replacement in any order

Example 6.3. *In the 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 on the balls 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 ball out of the machine has 59 possibilities, the second ball 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 balls could be drawn from the machine in any order! The sets of balls $\{1, 2, 3, 4, 5, 6\}$ 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 balls, which we know is 6!. So the number of possible sets of 6 balls 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

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{1}{\binom{59}{6}} \approx \frac{1}{45 \text{ million}} \approx 0.000\,000\,02.$$

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 6.4. *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 occur. We have

$$\begin{aligned}\mathbb{P}(A) = \frac{\binom{48}{9}}{\binom{52}{13}} &= \frac{\frac{48\times47\times\cdots\times41\times40}{9\times8\times\cdots\times2\times1}}{\frac{52\times51\times\cdots\times41\times40}{13\times12\times\cdots\times2\times1}} \\ &= \frac{48\times47\times\cdots\times41\times40}{52\times51\times\cdots\times41\times40} \times \frac{13\times12\times\cdots\times2\times1}{9\times8\times\cdots\times2\times1} \\ &= \frac{13\times12\times11\times10}{52\times51\times50\times49} \\ &\approx 0.0026,\end{aligned}$$

or about 1 in every 380 hands.

6.3 Birthday problem

Example 6.5. *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. If you haven't seen it before, you might like to guess what you think the answer might be. (We'll assume all days are equally likely for birthdays, and ignore the leap day 29 February.)

The sample space Ω 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 complementary event A^{c}. 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:

```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, and each of those 253 pairs is a potential match.

Summary

- Ordering n objects can be done in $n! = n^{\underline{n}} = n(n-1)\cdots 2 \cdot 1$ 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

You can download this problem sheet as a PDF file

This is Problem Sheet 2. This problem sheet covers material from Lectures 3 to 6. You should work through all the questions on this problem sheet in preparation for your tutorial in Week 4. The problem sheet contains two assessed questions, which are due in by **2pm on Monday 31 October**.

A: Short questions

A1. Suppose you toss a coin 4 times.

(a) What would you suggest for a sample space Ω **(i)** if you only care about the total number of heads; **(ii)** if you care about the result of each coin toss?

(b) For each of the cases in part (a), what is $|\Omega|$?

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. What is the value of the following expressions?

(a) $6!$

(b) 8^4

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

(d) $\dbinom{10}{4}$

A4. An urn contains 4 red balls and 6 blue balls. Two balls are drawn from the urn. What is the probability that both balls are 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) $\mathbb{P}(A^{\mathrm{c}}) = 1 - \mathbb{P}(A)$.

B2. 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 your result from part (a), the two-event addition rule, the distributive law, and the two-event addition rule again, prove the three-event form of the addition rule for unions:

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

B3. Suppose we pick a number at random from the set $\{1, 2, \dots, 2022\}$.

(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?

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

(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. By finding a relationship between B_k, A_{k-1} and A_k, or otherwise, show that

$$\mathbb{P}(B_k) = \frac{(k+1)^n - k^n}{10^n}.$$

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 31 October** at the beginning of Week 5. Submission is via Gradescope. Your work will be marked by your tutor and returned on Monday 7 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 Ω be a sample space with a probability measure $\mathbb{P}$, and let $A, B \subset \Omega$ be events. For each of the following statements, state whether the statement is true or false (that is, always true or sometimes false). If it is true, briefly justify the statement; if it is false, give a counterexample.

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

C2. An urn contains 15 balls: 4 red balls, 5 blue balls, and 6 green balls.

(a) If three balls are drawn *with* replacement, what is the probability that all three balls are the *same* colour?

(b) If three balls are drawn *without* replacement, what is the probability that all three balls are *different* colours?

Solutions to short questions

A1. (a) (i) $\{0, 1, \dots, 4\}$ (ii) $\{\mathrm{HHHH}, \mathrm{HHHT}, \mathrm{HHTH}, \dots, \mathrm{TTTT}\}$ (b) (i) 5 (ii) 16.
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) 720 (b) 4092 (c) 1680 (d) 210
A4. (a) 0.16 (b) 0.133

Chapter 7

Independence and conditional probability

7.1 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 would equally make 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 7.1. 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 in a question. But alternatively we might reason that A and B must be independent 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 – we could then use the independence assumption in calculations.

Example 7.1. *Consider rolling a dice. Let* $A = \{\textit{even number}\} = \{2, 4, 6\}$*, and let* $B = \{\textit{roll at least 4}\} = \{4, 5, 6\}$*. 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 7.2. *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

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

7.2 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%**

	Brown hair	**Blond hair**	**Total**
Blue eyes	10%	10%	**20%**
Total	**60%**	**40%**	**100%**

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

We know from this table that 20% of people have blue eyes. But suppose you already know that someone has blond hair: what *then* is the probability they 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 along with their blond hair, and the 30% of people who have brown eyes along with their blond hair. So of the 40% of blond-haired people, only one quarter of that 40% – which makes 10% – 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 7.2. 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 just 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.)

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.

Like with independence, we can use the definition of conditional probability in two ways. The first way is that if we know $\mathbb{P}(A \cap B)$ and $\mathbb{P}(A)$, then we can calculate the conditional probability of B given A as

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

Example 7.3. *With the dice roll again, what's the probability of rolling at least 4 given that you roll an even number?*

We previously calculated

$$\begin{aligned}\mathbb{P}(\text{even and at least } 4) &= \mathbb{P}(A \cap B) = \tfrac{1}{3} \\ \mathbb{P}(\text{even number}) &= \mathbb{P}(A) = \tfrac{1}{2}.\end{aligned}$$

Hence

$$\mathbb{P}(\text{even} \mid \text{at least } 4) = \mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \frac{\frac{1}{3}}{\frac{1}{2}} = \tfrac{2}{3}.$$

This is intuitively correct: of the three possibilities of rolling at least 4, $\{4, 5, 6\}$, two of those three are even, so the probability is $\frac{2}{3}$.

7.3 Chain rule

The second way to use the definition of conditional probability is that if we know $\mathbb{P}(A)$ and $\mathbb{P}(B \mid A)$, then we can calculate the event that both A and B occur as

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

This can be a particularly useful tool 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 conditional probability, given that A has already 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 7.1 (Chain rule)**.** *For events $A_1, A_2, \dots, A_n$, we have*

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

At each step, we need to calculate the probability of that step given all the previous steps being successful. We then multiply them all together.

Often questions that can be solved using the classical probability counting methods from Section 3 also be solved "step by step" using the chain rule. (It's a matter of personal taste which you prefer, but I find that chain rule methods are often simpler and more intuitive.)

Example 7.4. *Recall the Lotto problem from Example 6.3: 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 on our ticket that the first ball could match. 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, until $\mathbb{P}(A_6 \mid A_1 \cap \dots \cap A_5) = \frac{1}{54}$.

So, using the chain rule, we get

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

The answer we got before was

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

It's easy to see that this is the same answer. The structure of the answers shows how our previous classical probability method got the answer "all at once", while this new chain rule method gets the answer "one step at a time".

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 chain rule is

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

Chapter 8

Two theorems on conditional probability

Last time we met the conditional probability $\mathbb{P}(B \mid A)$ of one event B given another event A. In this lecture we will be looking two very useful theorems about conditional probability, called the **law of total probability** and **Bayes' theorem** – and they're particularly powerful when used together.

8.1 Law of total probability

Example 8.1. *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

$$\mathbb{P}(\text{roll } 5 \mid D_4) = 0 \qquad \mathbb{P}(\text{roll } 5 \mid D_6) = \tfrac{1}{6} \qquad \mathbb{P}(\text{roll } 5 \mid D_{10}) = \tfrac{1}{10}.$$

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 8.1. 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 8.1 (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. □

Example 8.1 continued. Returning to our dice example, we see that $\{D_4, D_6, D_{10}\}$ is indeed a partition, since my friend must choose exactly one of the three 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've already 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

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

8.2 Bayes' theorem

In this section, 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 8.2 (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. (Bayes' work was significantly edited by Richard Price, another minister–mathematician, and many historians 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 8.2. *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

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

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' theorem 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 this 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, using Bayes' theorem, to a "posterior" belief about the model. We will discuss Bayesian statistics much more in Week 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 some evidence B, then Bayes' theorem tells us how likely each hypothesis is given the evidence:

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

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

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

Putting these together, we get the following.

Theorem 8.3. *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 up the calculation into two separate parts rather than using this formula directly.

8.3 Diagnostic testing

Example 8.3. *Members of the public are tested for a certain rare 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 $\mathbb{P}(D^{\mathsf{c}}) = 0.98$;
- $\mathbb{P}(+ \mid D) = 0.95$;
- $\mathbb{P}(+ \mid D^{\mathsf{c}}) = 0.05$;
- we want to find $\mathbb{P}(D \mid +)$.

Importantly, D (you have the disease) and D^{c} (you don't) make up a partition. So Theorem 8.3 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

- The law of total probability says that if $A_1, A_2, \dots A_n$ is a partition of the sample space (that is, exactly one of them occurs), 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 9

Discrete random variables

9.1 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. 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, and the actual the two individual dice scores might be irrelevant. Let us wrote write X for the total score. This X is a sort of "numerical summary" of the experiment. Probabilists call such a numerical summary a **random variable**.

One we've defined this random variable X, it can be easier to work with X than with sample spaces and events. For example, if we want to know the probability that our two dice rolls add up to 5, it's more convenient to write

$$\mathbb{P}(X = 5)$$

rather than

$$\mathbb{P}(A) \quad \text{where} \quad A = \big\{(1,4),(2,3),(3,2),(4,1)\big\}.$$

The probability that $X = 5$ is $\mathbb{P}(X = 5) = \frac{4}{36} = \frac{1}{9}$. We might also be interested in other things about the total score X, like what the average total score over many pairs if dice rolls is.

The good news is that, once we have properly set up our random variable X, we can often then choose to ignore things like the sample space, the probability measure, and individual events.

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 the lower-case equivalent, like x, y, z.

This idea of a random variable as a numerical summary of an experiment is how we *think* about random variables when solving problems. On the other hand, as mathematicians, we also want to define carefully what a random variable is as a mathematical object. We'll discuss that now (but if you find the next couple of paragraphs difficult to follow, you won't miss much if you skip them).

In this formal mathematical view, our experiment of rolling two dice is represented by a sample space

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

of pairs $\omega = (\omega_1, \omega_2)$ of numbers from 1 to 6, where ω_1 represents the first dice roll and ω_2 the second dice roll. The random variable X is the score on the first dice plus the score on the second dice – that is,

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

In other words, X is a *function* which takes in a sample outcome $\omega \in \Omega$ and outputs a real number $X = X(\omega) = \omega_1 + \omega_2$.

Definition 9.1. 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 $\mathbb{P}(X = x)$ or $\mathbb{P}(X \in A)$ should be understood as representing more formal probabilities

$$\mathbb{P}(\{\omega : X(\omega) = x\}) \quad \text{or} \quad \mathbb{P}(\{\omega : X(\omega) \in A\}).$$

It's useful to have a notation for the values a random variable can take.

Definition 9.2. The set of values a random variable X can take is called its **range**, $\mathrm{Range}(X) = \{X(\omega) : \omega \in \Omega\}$.

So, for example, the range of the dice total X is $\mathrm{Range}(X) = \{2, 3, \dots, 12\}$, because those are the possible outcome from the sum of two dice rolls.

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

- **Discrete random variables** have a range that is a collection of discrete separate counts, so $\mathrm{Range}(X)$ is finite (like the dice total being an integer between 2 and 12) or countably infinite (like the positive integers). Discrete random variables can be used as models for "count data".
- **Continuous random variables** have a range that is a continuum of slowly varying measurements, so $\mathrm{Range}(X)$ is uncountably infinite (like the real numbers, the positive real numbers, or the interval $[0, 1]$). Continuous random variables can be used as models for "measurement data".

For this week and the two weeks after, we will look at discrete random variables; later in Lectures 15 to 17 we will look at continuous random variables.

9.2 Probability mass function

We now consider only discrete random variables X, where the range $\mathrm{Range}(X)$ is finite or countably infinite. To fully understand a discrete random variable X, we need only understand the probabilities $p(x) = \mathbb{P}(X = x)$. These are captured by the probability mass function.

Definition 9.3. 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 X is obvious from context, we'll just write $p(x)$ without the subscript.)

Once we have the PMF, then Axiom 3 tells us that for any set A, we have

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

(Recall that the symbol $\in$ means "is an element of", or just "is in" for short.) So the probability that X is in some set A can be found by simply adding up $p(x)$ for all the values x in A. Thus the PMF $p(x)$ is the only thing we need to know.

Example 9.1. 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

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

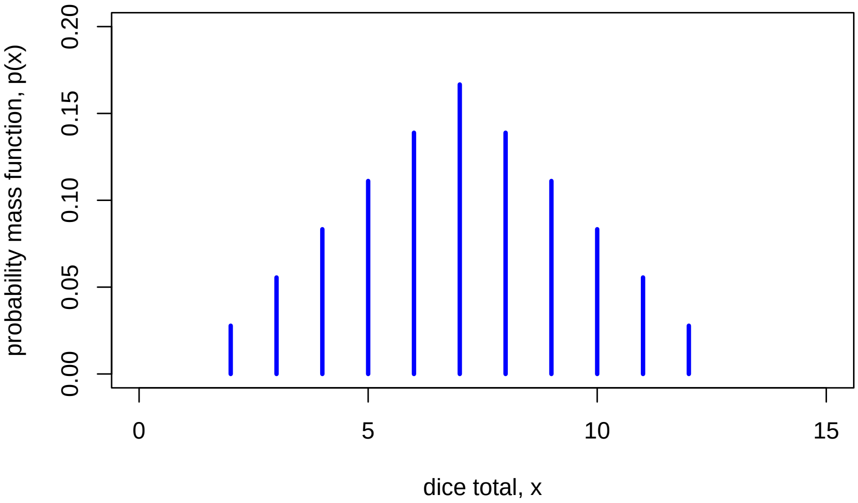
We could alternatively think of the same random variable X as representing the result of an experiment, where $X = 1$ represents a success, with probability $p_X(1) = p$, and $X = 0$ represents a failure, with probability $p_X(0) = 1 - p$.

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 $X \sim \text{Bern}(p)$ for short.

Example 9.2. 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	8	9	10	11	12
$p(x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$



Note that since $p(x) = \mathbb{P}(X = x)$ is a probability, it must be greater than or equal to 0, by Axiom 1. Further, if we add up $p(x)$ we get

$$\sum_x p(x) = \sum_{x \in \text{Range}(X)} \mathbb{P}(X = x) = \mathbb{P}(X \in \text{Range}(X)) = \mathbb{P}(\Omega) = 1,$$

by Axiom 2. Hence we have the following:

Theorem 9.1. *Let X be a discrete random variable, and let p_X be its PMF. Then*

- $p_X(x) \geq 0$ *for all* $x \in \text{Range}(X)$*;*
- $\displaystyle\sum_{x \in \text{Range}(X)} p_X(x) = 1.$

9.3 Cumulative distribution function

Sometimes it is useful to know the probability a random variable X is less or equal to 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 9.3. Let $X \sim \text{Bern}(p)$ be a Bernoulli random variable with success probability p. It's impossible for the outcome to be less than 0; it is at least 0 but strictly less than 1 only if it equals 0, which happens with probability

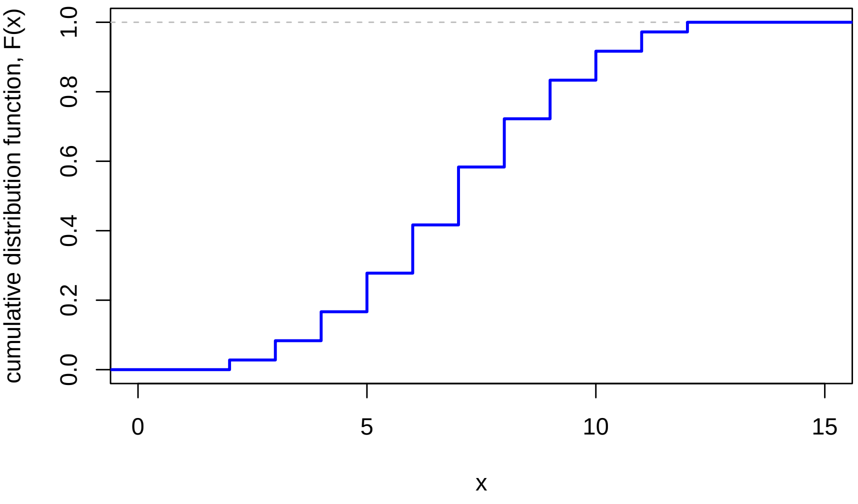
$p(0) = 1 - p$; and it is certain to at most 1. So its CDF F is

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

Example 9.4. If X is the sum of two dice rolls, then the CDF F is given by adding up the PMF. So, for example:

- If $x < 2$, then $F(x) = 0$, because it's not possible to have $X \le x < 2$.
- If $2 \le x < 3$, then $F(x) = p(2) = \frac{1}{36}$, because $X = 2$ is the only outcome with $X \le x < 3$.
- If $3 \le x < 4$, then $F(x) = p(2) + p(3) = \frac{1}{36} + \frac{2}{35} = \frac{3}{36}$, as $X = 2$ or 3 are the only outcomes with $X \le x$.
- ...
- If $x \ge 12$, then $F(x) = 1$, because we always have $X \le 12 \le x$.

$x \in$	$(-\infty, 2)$	$[2, 3)$	$[3, 4)$	$[4, 5)$	$\cdots$	$[11, 12)$	$[12, \infty)$
$F(x)$	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 $2, 3, \dots, 12$, finally ending up at 1.

For any random variable X with CDF F,

- if x is smaller than everything in the Range(X), then $F(x) = 0$, because X cannot be that small;

- if x is greater than everything in the Range(X), then $F(x) = 1$, because X cannot be any bigger than that;
- $F(x)$ is increasing in x, because the probability you are less than x gets bigger as x gets bigger.

Summary

- A random variable is a numerical summary of a random experiment.
- The probability mass function (PMF) is $p_X(x) = \mathbb{P}(X = x)$.
- The cumulative distribution function (CDF) is $F_X(x) = \mathbb{P}(X \leq x)$.
- A Bernoulli random variable is 0 with probability $1 - p$ and 1 with probability p.

Chapter 10

Expectation and variance

We continue our study of discrete random variables. Recall that the PMF p_X of a discrete random variable X is $p_X(x) = \mathbb{P}(X = x)$.

10.1 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 the "central" value of the random variable is. This average is called the "expectation".

Definition 10.1. 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_X 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. It's only this second formula we actually use when calculating expectations.

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 in the module that the the expectation can be interpreted as a sort of "long-run mean outcome".

Example 10.1. *Let $X \sim$ Bern(p) be a Bernoulli trial with success probability p. What is the expectation $\mathbb{E}X$?*

We know that the PMF is $p(0) = 1 - p$ and $p(1) = p$. So, 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 10.2. *What is the expected value of the sum of two dice rolls?*

When X is the total of two dice rolls, we found the PMF of X last time. The expectation is

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

10.2 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...)

It is possible, although sometimes a bit tricky, to work out the whole PMF of these new random variables that are functions of X – and indeed you may learn how to do this if you take more probability or statistics modules next year. Here, we will stick to the easier problem of just calculating the expectation of the new random variables.

Theorem 10.1 (Law of the unconscious statistician)**.** *Let X be a random variable, and let $Y = g(X)$ be another random variable that is a function g 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. (Non-examinable) The PMF of y can be given in terms of the PMF of x – we just need to add up all the xs that lead to the same y. That is,

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

(Remember that the colon : here means "such that", so this is a sum over all the x such that $g(x) = y$.)

Using this, and from the definition of expectation, we have

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

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

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

Theorem 10.2 (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

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

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

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

10.3 Variance

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

Definition 10.2. 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 here means the expectation of $(X - \mu)^2$; and *not* $\mathbb{E}(X - \mu)$ then squared, which would be $0^2 = 0$.)

Note that $(X - \mu)^2 \geq 0$ is a square, so always non-negative, and hence the variance $\mathrm{Var}(X) \geq 0$ is always non-negative also. Sometimes we call the square-root of the variance the **standard deviation**.

It may not surprise you, if you remember Lecture 1 that to go along with this "definitional formula" for the variance, we also have a "computational formula", which can sometimes be more convenient.

Theorem 10.3. *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" with the sample variance previously). We get

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

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

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

as required. □

(A brief optional note for pedants: Writing $\mathbb{E}(X^2 - 2\mu X) = \mathbb{E}X^2 - 2\mu X$ is not, strictly speaking, justified by the result that above we called "linearity of expectation, 1". However, you can check that it easily follows from the law of the unconscious statistician, and we will also later see a result we call "linearity of expectation, 2", of which it is a special case.)

Example 10.3. Let $X \sim \mathrm{Bern}(p)$ be a Bernoulli trial, and recall that $\mathbb{E}X = p$. Using the definitional formula, we have

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

Alternatively, using the computational formula, we have

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

Example 10.4. For the total of two dice, using the computational formula, we have

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

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

Theorem 10.4. *Let X be a random variable. Then*

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

You will prove this on the problem sheet.

Summary

- The expectation of a random variable X is $\mathbb{E}X = \sum_x x\, p_X(x)$.
- The variance of a random variable X 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

Full **solutions to non-assessed questions** are now available.

The **mid-semester check-in** survey is still open.

You can download this problem sheet as a PDF file

This is Problem Sheet 3. This problem sheet covers material from Lectures 7 to 10. You should work through all the questions on this problem sheet in preparation for your tutorial in Week 6. The problem sheet contains two assessed questions, which are due in by **Monday 14 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. Consider rolling two dice. 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. You may assume the dice rolls are independent; so, in particular, events A and B are independent.

(a) Are A and C independent?

(b) Are B and C independent?

(c) Is it true that $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\,\mathbb{P}(B)\,\mathbb{P}(C)$?

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	0.1

This random variable has $\mathbb{E}X = 4.3$. Find the values of a and b.

A5. A 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. The variance of the daily maximum temperature measured in degrees Celsius is 10.4.

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

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

B: Long questions

B1. Suppose A and B are independent events. Show that A and B^{c} are also independent events.

B2. You are dealt a hand of 13 cards from a 52-card deck. Let $E_{\mathrm{A}}, E_{\mathrm{K}}, E_{\mathrm{Q}}, E_{\mathrm{J}}$ respectively be the events that your hand contains the Ace, King, Queen and Jack of Spades.

(a) What is $\mathbb{P}(E_{\mathrm{A}})$, the probability that your hand contains the Ace of Spades?

(b) Explain why $\mathbb{P}(E_{\mathrm{K}} \mid E_{\mathrm{A}}) = \frac{12}{51}$.

(c) Using the chain rule, calculate the probability that your hand contains all four of the Ace, King, Queen and Jack of Spades.

(d) Check that your answer agrees with the answer we found by classical probability methods in Example 6.4 in Lecture 6. Which method do you prefer?

B3. Soldiers are asked about their use of illegal drugs, using a so-called "randomised survey". 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) What are some advantages or disadvantages of performing the experiment this way?

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

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

B4. A random variable X_n is said to follow the *discrete uniform distribution* on $\{1, 2, \dots, n\}$ if each of the n values in that set $\{1, 2, \dots, n\}$ is equally likely.

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

(b) Find the variance of X_n.

(c) Let Y be a discrete uniform distribution on $b - a + 1$ values $\{a, a+1, a+2, \dots, b-1, b\}$, 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.

*[**Note:** "$b - a + 1$ values" is correct, but this was wrong earlier.]*

You may use without proof the standard results

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

B5. A gambling game works as follows. You keep tossing a fair coin until you first get a Head. If the first Head comes up 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 observation that, despite the expected winnings from this game being infinite, few people would be prepared to play this game for, say, £100, and almost no one 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 14 November** at the beginning of Week 7. Submission is via Gradescope. Your work will be marked by your tutor and returned on Monday 21 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? Explain your solution fully.

C2. Let X be a random variable.

(a) Let $Y = aX$ be another random variable. What is $\mathbb{E}Y$, in terms of $\mu = \mathbb{E}X$?

(b) Using part (a), show that $\operatorname{Var}(aX) = a^2 \operatorname{Var}(X)$.

(c) Prove that $\operatorname{Var}(X + b) = \operatorname{Var}(X)$.

Solutions to short questions

A1. (c) and (e) are independent.
A2. (a) Yes (b) Yes (c) No
A3. 0.45 and 0.5725
A4. $a = 9, b = 0.5$
A5 (a) 66.2 °F (b) 33.7 $°\mathrm{F}^2$

Chapter 11

Binomial and geometric distributions

Last week, 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. This week, we will look at a number of useful discrete distributions.

11.1 Binomial distribution

One family of distributions we have already seen is the Bernoulli trial $\mathrm{Bern}(p)$, 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 11.1. *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

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

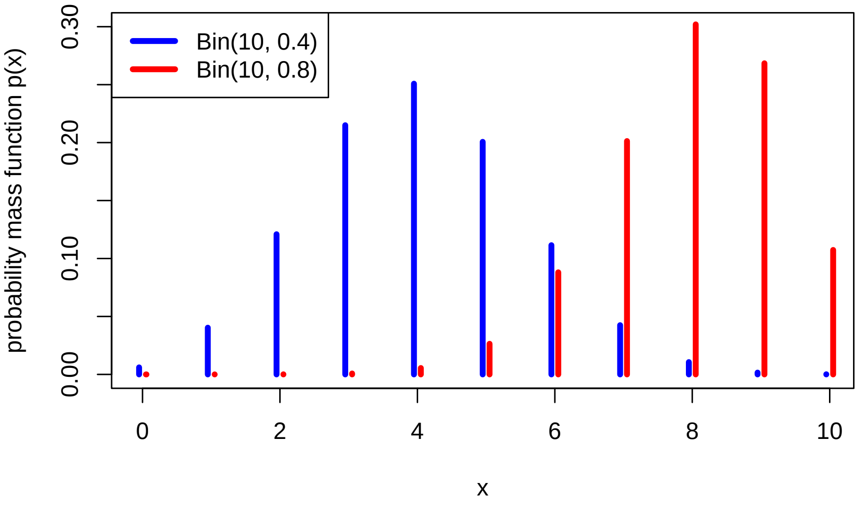
This is a special case of the binomial distribution.

Definition 11.1. 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 $X \sim \mathrm{Bin}(n, p)$.

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 $\mathrm{Bin}(10, 0.7)$.



Example 11.2. *Let* $X \sim \mathrm{Bin}(8, 0.2)$*. What is (a)* $\mathbb{P}(X = 3)$*? (b)* $\mathbb{P}(X \geq 2)$*?*

For (a), we have from the definition

$$\mathbb{P}(X = 3) = \binom{8}{3} 0.2^3 (1-0.2)^{8-3} = 56 \times 0.2^3 \times 0.8^5 = 0.147.$$

For (b), this is an "at least" question, so it's more convenient to look at the complementary event, $\mathbb{P}(X < 2)$. So

$$\begin{aligned}\mathbb{P}(X \geq 2) &= 1 - \mathbb{P}(X < 2) \\ &= 1 - \mathbb{P}(X = 0) - \mathbb{P}(X = 1) \\ &= 1 - 0.8^8 - 8 \times 0.2 \times 0.8^7 \\ &= 1 - 0.168 - 0.336 \\ &= 0.497.\end{aligned}$$

What about the expectation and variance of a binomial random variable?

Theorem 11.1. *Let $X \sim Bin(n, p)$. Then*

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

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

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

which is a bit tricky to calculate, but not fundamentally difficult mathematics. However, in next section we will see an easier way, so we'll reserve the proof until then instead.

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

$$\begin{aligned} \mathbb{E}X &= 10 \times 0.7 = 7 \\ \operatorname{Var}(X) &= 10 \times 0.7 \times 0.3 = 2.1 \end{aligned}$$

11.2 Geometric distribution

Example 11.3. *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

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

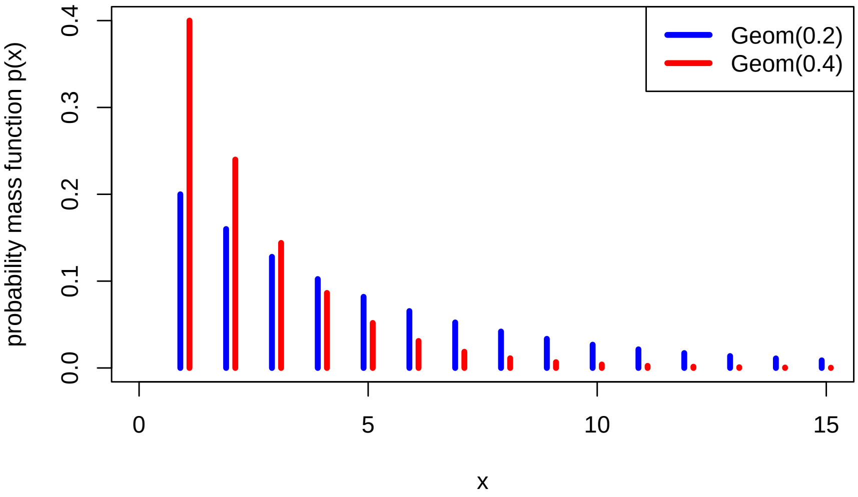
This is a special case of the geometric distribution.

Definition 11.2. 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 $X \sim \operatorname{Geom}(p)$.

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



Example 11.4. *Let* $X \sim \mathrm{Geom}(0.4)$*. What is (a)* $\mathbb{P}(X=3)$*? (b)* $\mathbb{P}(X \geq 3)$*.*

For part (a), we have

$$\mathbb{P}(X=3) = (1-0.4)^2 \times 0.4 = 0.144.$$

For part (b), we have

$$\mathbb{P}(X \geq 3) = 1-\mathbb{P}(X=1)-\mathbb{P}(X=2) = 1-0.4-(1-0.4)\times 0.4 = 1-0.64 = 0.36.$$

Theorem 11.2. *Let* $X \sim Geom(p)$*. Then*

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

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

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

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

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

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

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

□

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 before success" definition, as we will discuss in a later R worksheet.

11.3 Distributions as models for data

Families of distributions – like the Bernoulli, binomial and geometric distributions we have seen so far in this module – are very useful for models in statistics. This idea is developed Bayesian statistics we will discuss in Lectures 19 and 20 of this module, and is 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; and both n and p for the binomial distribution. (In the next lecture we will see another discrete distribution, the Poisson distribution, and 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 one of the distributions within a family, then use data to estimate the value of that parameter.

For example:

- When testing the bias of a coin, you might assume, counting Heads as 1 and Tails as 0, that the outcome of each test is Bernoulli distributed with parameter p, but where the value of the Heads probability p is unknown. You could then toss the coin many times and use this data to estimate p.
- The number of years between severe summer floods in a tropical climate could be modelled as geometrically distributed where the flood risk parameter p is unknown. By look at the gaps between severe floods in historical data, a statistician could try to estimate p.
- 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, the number of students assigned to the group, but p, the "turning-up probability" 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:

- **Bayesian statistics:** Here, one starts with a subjective "prior" distribution for the parameters, which represents one's personal belief about which possible values for that parameter are more or less likely before conducting any experiment. After the experiment, 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 Lecture 19 of this module.

- **Frequentist statistics:** Frequentist statistics does not involve any subjective prior views. Rather, frequentism is about assessing the extent to which the data is consistent with certain hypotheses. For example, one might try to find the value for the parameter that would seem "most consistent" with the data, and use that as an "estimate" of the parameter: "My best guess of the Heads probability p of the coin is $p = 0.53$." Alternatively, one might find a range of values for the parameter that are all at least somewhat consistent with the data: "I am confident the value of the Heads probability lies in the range $0.49 \leq p \leq 0.57$, as these values are all consistent with the data." Third, one could text if a specific hypothesis is consistent with the data or not: "The data is consistent with the hypothesis that $p = 0.50$, whih would mean the coin is fair, so I have no strong evidence for disbelieving that." The frequentist approach will pursued in detail in MATH1712.

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)	$\{1, 2, \dots\}$	$(1-p)^{x-1}p$	$\dfrac{1}{p}$	$\dfrac{1-p}{p^2}$

Chapter 12

Poisson distribution

We have seen three important families of discrete random variables: the Bernoulli, binomial, and geometric distributions. We now look at our fourth and final discrete distribution: the Poisson distribution.

12.1 Definition and properties

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 own horse between 1875 and 1894".) We'll explain why the Poisson distribution is a good model for this in the next subsection.

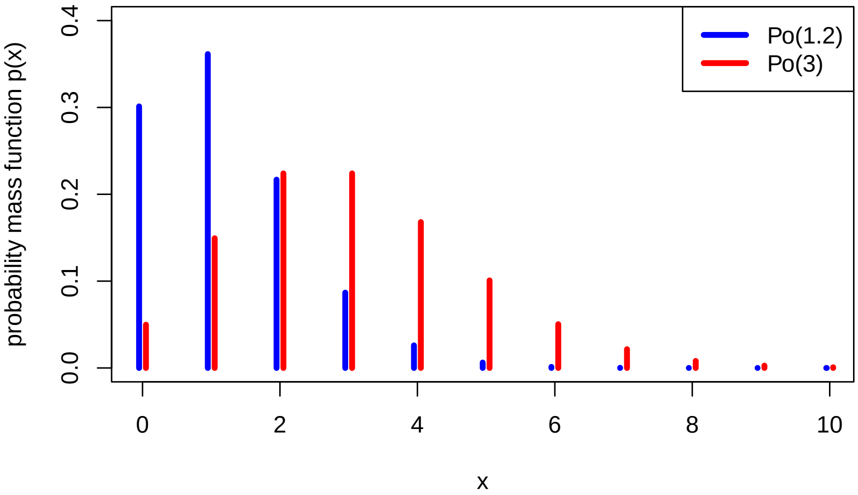
Definition 12.1. Let X be a discrete random variable with range $\{0, 1, 2, \dots\}$ and PMF

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

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

Here, λ is a lower-case Greek letter "lambda". I should also 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 12.1. *I receive emails from students at the rate of $\lambda = 3$ per hour, modelled as a Poisson distribution. What is the probability I get (a) two email in an hour, (b) no email in an hour?*

The number of emails per hour is $X \sim \mathrm{Po}(3)$.

For (a), we have

$$\mathbb{P}(X = 3) = p(3) = \mathrm{e}^{-3}\frac{3^2}{2!} = \frac{9}{4}\mathrm{e}^{-3} = 0.224.$$

For part (b), and remembering that $0! = 1$, we have

$$\mathbb{P}(X = 3) = p(0) = \mathrm{e}^{-3}\frac{3^0}{0!} = \mathrm{e}^{-3} = 0.050.$$

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

1. *$p(x)$ is indeed a PMF, in that $\displaystyle\sum_{x=0}^{\infty} p(x) = 1$.*
2. *$\mathbb{E}X = \lambda$,*
3. *$\mathrm{Var}(X) = \lambda$.*

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,

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

For part 1, 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 part 2, 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. □

At this point in the lecture, we took a break to fill in the **mid-semester check-in survey**. This is open for the rest of the week and is anonymous. Written comments in answer to the last question are particular useful – I will read them all and report back next week on changes I am making to the module in response to your comments.

12.2 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 some 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 must 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 12.2. *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 12.2. *A lecturer teaches a module with $n = 100$ students, and estimates 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 have an "at least" event, so 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

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

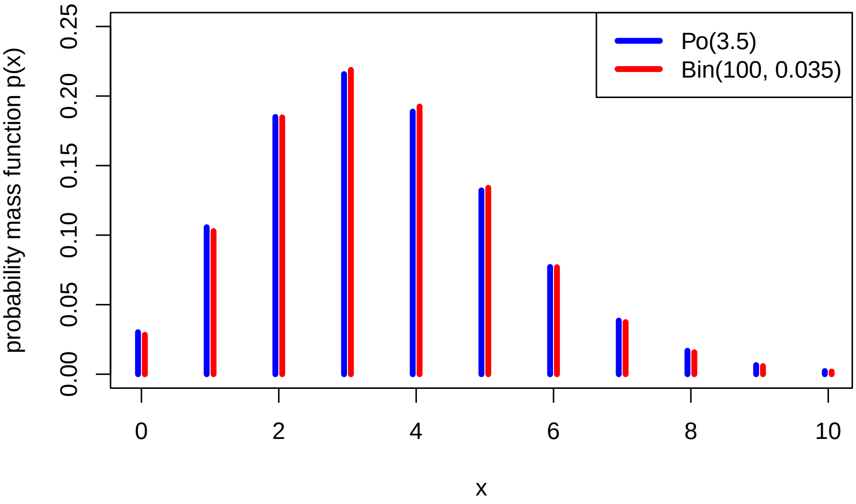
which is very close to the exact answer above of 0.134.

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 12.2 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 $n \to \infty$,

$$p_X(x) = \binom{n}{x}\left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \to \mathrm{e}^{-\lambda}\frac{\lambda^x}{x!} = p_Y(x).$$

Let's try. The left-hand side is, by some simple rearrangements,

$$\begin{aligned}
&\binom{n}{x}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x} \\
&\qquad = \frac{n(n-1)\cdots(n-x+1)}{x!}\frac{\lambda^x}{n^x}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-x} \\
&\qquad = \frac{\lambda^x}{x!}\frac{n(n-1)\cdots(n-x+1)}{n^x}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-x} \\
&\qquad = \frac{\lambda^x}{x!}\frac{n}{n}\frac{n-1}{n}\cdots\frac{n-x+1}{n}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-x} \\
&\qquad = \frac{\lambda^x}{x!}1\left(1-\frac{1}{n}\right)\cdots\left(1-\frac{x-1}{n}\right)\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-x}.
\end{aligned}$$

Now let's take each of the terms in turn. First $\lambda^x/x!$ looks very promising, and can stay. Second, each of the terms $1, 1-1/n, \dots, 1-(x-1)/n$ tend to 1 as $n \to \infty$. Third,

$$\left(1-\frac{\lambda}{n}\right)^n \to \mathrm{e}^{-\lambda};$$

this is from the standard "compound interest" 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. □

12.3 Poisson process

Suppose an insurance company's call centre is open 10 hours a day, 5 days a week. The call centre receives a "large claim" – a claim in excess of £100,000 – on average 0.2 times per hour. It seems reasonable, therefore, to model the number of large claims in an hour as a $\mathrm{Po}(\lambda)$ distribution with $\lambda = 0.2$.

How should we model the nuber of large claims in a day? If the call centre receives $\lambda = 0.2$ claims per hour, on average, then it receives $10\lambda = 2$ claims per 10 hours, or one day. It seems reasonable to model this with a $\mathrm{Po}(10\lambda)$ distribution.

Further, the number of claims on one day and on the next day seems like they should be independent.

The two ideas in the model above lead to what is called the "Poisson process".

Definition 12.2. A random set of arrivals is said to follow a **Poisson process** with rate λ if

1. The number of arrivals in a time period of length t is $\mathrm{Po}(\lambda t)$.

2. The number of arrivals in two non-overlapping time periods are independent.

Example 12.3. *Suppose the number of large claims, as discussed above, is modelled as a Poisson process with rate $\lambda = 0.2$ claims per hour. What is the probability the call centre receives at least one large claim every day this week?*

The number of large claims in one 10-hour day is $X \sim \mathrm{Po}(10\lambda) = \mathrm{Po}(2)$. So the probability of getting at least one claim in a day is

$$\mathbb{P}(X \geq 1) = 1 - \mathbb{P}(X = 0) = 1 - \mathrm{e}^{-2} = 0.865.$$

Because the number of claims in each of the five days this week are independent, the probability of getting at least one claim in all five days is

$$\mathbb{P}(X \geq 1)^5 = 0.865^5 = 0.483.$$

This is just a brief taster of the Poisson process. The Poisson process is studied in much more detail in the second-year module MATH2750 Introduction to Markov Processes.

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)	$\{1,2,\dots\}$	$(1-p)^{x-1}p$	$\dfrac{1}{p}$	$\dfrac{1-p}{p^2}$
Poisson: Po(λ)	$\{0,1,\dots\}$	$\mathrm{e}^{-\lambda}\dfrac{\lambda^x}{x!}$	λ	λ

Chapter 13

Multiple random variables

13.1 Joint distributions

In previous sections, we have looked at single discrete random variables in isolation. In the lecture and the next, 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 just one worked example, we have $p_{X,Y}(2,2) = \frac{1}{8}$, because the only way to have $X = 2$ (two Heads in the first two coin tosses) and $Y = 2$ (two Heads in the first *three* coin tosses) is to toss Head, Head, Tail, with probability $(\frac{1}{2})^3 = \frac{1}{8}$.

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.)

In summary, we have learned the following:

- A joint PMF of two discrete random variables X and Y is
$$p_{X,Y}(x,y) = \mathbb{P}(X = x \text{ and } Y = y).$$
- We can get the marginal distributions $p_X(x) = \mathbb{P}(X = x)$ and $p_Y(y) = \mathbb{P}(Y = y)$ by summing over the other variable:
$$p_X(x) = \sum_y p_{X,Y}(x,y) \qquad p_Y(y) = \sum_x p_{X,Y}(x,y).$$

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: $\sum\limits_{x,y} p_{X,Y}(x,y) = 1$.

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:
$$p_{X_1}(x) = \sum_{x_2,\dots,x_n} p_{X_1,X_2,\dots,X_n}(x, x_2, \dots, x_n).$$

13.2 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 13.1. We say two discrete random variables are **independent** if, for all x and y, the events $\{X = x\}$ and $\{Y = y\}$ are independent; that is, if
$$\mathbb{P}(X = x \text{ and } Y = y) = \mathbb{P}(X = x)\,\mathbb{P}(Y = y).$$

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 variables $\mathbf{X} = (X_1, X_2, \dots)$ are independent if
$$p_{\mathbf{X}}(\mathbf{x}) = p_{X_1}(x_1) \times p_{X_2}(x_2) \times \cdots = \prod_i p_X(x_i).$$

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 $p_Y(0) = \frac{1}{8}$, so $p_{X,Y}(0,0) \neq p_X(0)\, p_Y(0)$.

An important scenario in probability theory and statistics is that of **independent and identically distributed** (or **IID**) random variables. IID random variables represent the same experiment being 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 $\mathbf{X} = (X_1, X_2, \dots)$ are IID random variables with a common PMF p_X, say, if

$$p_{\mathbf{X}}(\mathbf{x}) = p_X(x_1) \times p_X(x_2) \times \cdots = \prod_i p_X(x_i).$$

Example 13.1. *Let $X_1, X_2, \dots, X_{20}$ be IID random variables following a binomial distribution with rate $n = 10$ and $p = 0.3$. What is the probability that all 20 of the the X_i are nonzero?*

Because the X_i are identically distributed, the probability that any one of them is nonzero is

$$\mathbb{P}(X_1 > 0) = 1 - \mathbb{P}(X_1 = 0) = 1 - (1 - 0.3)^{10} = 0.972.$$

Then, because the X_i independent, the probability that they are all nonzero is

$$\mathbb{P}(X_1 > 0)^{20} = 0.972^{20} = 0.564.$$

13.3 Conditional distributions

For probabilities of 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

$$\mathbb{P}(Y = y \mid X = x) = \frac{\mathbb{P}(X = x \text{ and } Y = y)}{\mathbb{P}(X = x)}.$$

We call the distribution of this the **conditional distribution**.

Definition 13.2. 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 $p_X(1)$ 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 by $p_Y(2)$ so that the column sums to 1, That is,

$$\begin{aligned}
p_{X|Y}(0 \mid 2) &= \frac{p_{X,Y}(0,2)}{p_Y(2)} = \frac{0}{\frac{3}{8}} = 0, \\
p_{X|Y}(1 \mid 2) &= \frac{p_{X,Y}(1,2)}{p_Y(2)} = \frac{\frac{1}{4}}{\frac{3}{8}} = \tfrac{2}{3}, \\
p_{X|Y}(2 \mid 2) &= \frac{p_{X,Y}(2,2)}{p_Y(2)} = \frac{\frac{1}{8}}{\frac{3}{8}} = \tfrac{1}{3}.
\end{aligned}$$

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 **Bayes' theorem** for random variables:

$$\mathbb{P}(X = x \mid Y = y) = \frac{\mathbb{P}(X = x)\,\mathbb{P}(Y = y \mid X = x)}{\mathbb{P}(Y = y)},$$

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

$$\frac{p_X(1)\,p_{Y|X}(2 \mid 1)}{p_Y(2)} = \frac{\frac{1}{2} \times \frac{1}{2}}{\frac{3}{8}} = \tfrac{2}{3}.$$

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.

Summary

- For two random variables, the joint PMF $p_{X,Y}$, marginal PMF p_X, and conditional PMF $p_{Y|X}$ are

$$\begin{aligned} p_{X,Y}(x,y) &= \mathbb{P}(X = x \text{ and } Y = y) \\ p_X(x) &= \mathbb{P}(X = x) = \sum_y p_{X,Y}(x,y) \\ p_{Y|X}(y \mid x) &= \mathbb{P}(Y = y \mid X = x) = \frac{p_{X,Y}(x,y)}{p_X(x)} \end{aligned}$$

- Two random variables are independent if $p_{X,Y}(x,y) = p_X(x)\, p_Y(y)$.
- Bayes' theorem:

$$p_{X|Y}(x \mid y) = \frac{p_X(x)\, p_{Y|X}(y \mid x)}{p_Y(y)}.$$

Chapter 14

Expectation and covariance

14.1 Expectation of 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. 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 14.1. *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 10.2) 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 14.1. *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.

14.2 Covariance

If we are interested at how two random variables vary together, we need to look at the covariance.

Definition 14.1. Let X and Y be two random variables with expectations $\mathbb{E}X = \mu_X$ and $\mathbb{E}Y = \mu_Y$ respectively. Then their **covariance** is

$$\operatorname{Cov}(X,Y) = \mathbb{E}(X-\mu_X)(Y-\mu_Y).$$

In the least surprising result of this whole module, we also have a computational formula to go along with this definitional formula.

Theorem 14.2. *Let X and Y be two random variables with expectations μ_X and μ_Y respectively. Then their covariance can also be calculated as*

$$\operatorname{Cov}(X,Y) = \mathbb{E}XY - \mu_X\,\mu_Y.$$

Proof. Exactly as we've done many times before, we have

$$\begin{aligned}\operatorname{Cov}(X,Y) &= \mathbb{E}(X-\mu_X)(Y-\mu_Y) \\ &= \mathbb{E}(XY - X\,\mu_Y - \mu_X\,Y + \mu_X\,\mu_Y) \\ &= \mathbb{E}XY - \mu_Y\,\mathbb{E}X - \mu_X\,\mathbb{E}Y + \mu_X\,\mu_Y \\ &= \mathbb{E}XY - \mu_X\,\mu_Y - \mu_X\,\mu_Y + \mu_X\,\mu_Y \\ &= \mathbb{E}XY - \mu_X\,\mu_Y,\end{aligned}$$

and we're done. □

Example 14.2. We continue with our coin-tossing example from the last lecture, where X is the number of Heads in the first two coin tosses and Y the number of Heads in the first three coin tosses.

We know that $X \sim \mathrm{Bin}(2, \frac{1}{2})$, so $\mu_X = 1$, and $Y \sim \mathrm{Bin}(3, \frac{1}{2})$, so $\mu_Y = 1.5$. To find the covariance using the computational formula, we also need $\mathbb{E}XY$, which is

$$\begin{aligned}\mathbb{E}XY &= \sum_{x,y} xy\,p_{X,Y}(x,y) \\ &= 0\times 0\times p_{X,Y}(0,0) + 0\times 1\times p_{X,Y}(0,1) + \cdots + 2\times 3\times p_{X,Y}(2,3) \\ &= 0\times \tfrac{1}{8} + 0\times\tfrac{1}{8} + \cdots + 6\times\tfrac{1}{8} \\ &= 2.\end{aligned}$$

Hence the covariance is

$$\operatorname{Cov}(X,Y) = \mathbb{E}XY - \mu_X\mu_Y = 2 - 1\times 1.5 = 0.5.$$

A very important fact is the following.

Theorem 14.3. *If X and Y are independent, then* $\operatorname{Cov}(X,Y) = 0$.

Be careful not to get this the wrong way around: if $\operatorname{Cov}(X,Y) = 0$ it doesn't necessarily mean that X and Y are independent.

To use the “contrapositive” (which is allowed!), in our example, we have $\operatorname{Cov}(X,Y) \neq 0$, which means that X and Y are not independent – confirming what we already knew.

Proof. Recall from Theorem 14.1 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

$$\operatorname{Cov}(X,Y) = \mathbb{E}XY - \mu_X\,\mu_Y = \mu_X\,\mu_Y - \mu_X\,\mu_Y = 0,$$

and we are done. □

Here are some more important properties of the covariance.

Theorem 14.4. *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

$$\begin{aligned}\mathrm{Cov}(X + Y, Z) &= \mathbb{E}((X + Y) - (\mu_X + \mu_Y))(Z - \mu_Z)\\ &= \mathbb{E}((X - \mu_X) + (Y - \mu_Y))(Z - \mu_Z)\\ &= \mathbb{E}((X - \mu_X)(Z - \mu_Z) + (Y - \mu_Y)(Z - \mu_Z))\\ &= \mathbb{E}(X - \mu_X)(Z - \mu_Z) + \mathbb{E}(Y - \mu_Y)(Z - \mu_Z)\\ &= \mathrm{Cov}(X, Z) + \mathrm{Cov}(Y, Z),\end{aligned}$$

as required. □

We could calculate the covariance in our coin-tossing example a different way, by noting that $Y = X + Z$, where $Z \sim \mathrm{Bern}(\frac{1}{2})$ represents the third coin toss and is independent of X. Then we have

$$\mathrm{Cov}(X, Y) = \mathrm{Cov}(X, X+Z) = \mathrm{Cov}(X, X)+\mathrm{Cov}(X, Z) == \mathrm{Var}(X)+0 = \mathrm{Var}(X),$$

where we used $\mathrm{Cov}(X, Z) = 0$ since X and Z are independent. We already know that $\mathrm{Var}(X) = 2 \times \frac{1}{2} \times (1 - \frac{1}{2}) = \frac{1}{2}$ because $X \sim \mathrm{Bin}(2, \frac{1}{2})$.. So $\mathrm{Cov}(X, Y) = \frac{1}{2}$, matching our previous calculation.

Now that we know some facts about the covariance, we can calculate the variance of a sum.

Theorem 14.5. *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

$$\begin{aligned}\operatorname{Var}(X+Y) &= \mathbb{E}\big((X+Y) - (\mu_X + \mu_Y)\big)^2 \\ &= \mathbb{E}\big((X-\mu_X) + (Y-\mu_Y)\big)^2 \\ &= \mathbb{E}\big((X-\mu_X)^2 + 2(X-\mu_X)(Y-\mu_Y) + (Y-\mu_Y)^2\big) \\ &= \mathbb{E}(X-\mu_X)^2 + 2\mathbb{E}(X-\mu_X)(Y-\mu_Y) + \mathbb{E}(Y-\mu_Y)^2 \\ &= \operatorname{Var}(X) + 2\operatorname{Cov}(X,Y) + \operatorname{Var}(Y),\end{aligned}$$

where we used the linearity of expectation.

For the second part, recall that is X and Y are independent, then $\operatorname{Cov}(X,Y) = 0$. □

14.3 Correlation

It can sometimes be useful to "normalise" the covariance, by dividing through by the individual standard deviations. This gives a measurement of the linear relationship between two random variables.

Definition 14.2. Let X and Y be two random variables. Then the **correlation** between X and Y is

$$\operatorname{Corr}(X,Y) = \frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}}.$$

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 $\operatorname{Cov}(X,Y) = 0$. Hence it follows that if X and Y are independent, then $\operatorname{Corr}(X,Y) = 0$ also.

Example 14.3. For the coin-tossing again, we have

$$\operatorname{Corr}(X,Y) = \frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}} = \frac{\frac{1}{2}}{\sqrt{\frac{1}{2}\times\frac{3}{4}}} = \sqrt{\tfrac{2}{3}} = 0.816.$$

Summary

- $\mathbb{E}(X+Y) = \mathbb{E}X + \mathbb{E}Y$
- The covariance is $\operatorname{Cov}(X,Y) = \mathbb{E}(X-\mu_X)(Y-\mu_Y) = \mathbb{E}XY - \mu_X\,\mu_Y$.
- $\operatorname{Var}(X+Y) = \operatorname{Var}(X) + 2\operatorname{Cov}(X,Y) + \operatorname{Var}(Y)$; or if X and Y are independent, then $\operatorname{Var}(X+Y) = \operatorname{Var}(X) + \operatorname{Var}(Y)$.

Problem Sheet 4

This is Problem Sheet 4. This problem sheet covers Lectures 11 to 14. You should work through all the questions on this problem sheet in preparation for ~~your~~ **the online** tutorial in Week 8. The problem sheet contains two assessed questions, which are due in by **2pm on** ~~Monday 28~~ **Tuesday 29 November**.

A: Short questions

A1. Let $X \sim \mathrm{Bin}(20, 0.4)$. Calculate

(a) $\mathbb{P}(X = 8)$

(b) $\mathbb{P}(8 \leq X \leq 11)$

(c) $\mathbb{E}X$

A2. Let $X \sim \mathrm{Geom}(0.2)$. Calculate

(a) $\mathbb{P}(X = 2)$

(b) $\mathbb{P}(X \geq 3)$

(c) $\mathrm{Var}(X)$

A3. Let $X \sim \mathrm{Po}(2.5)$. 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?

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 $\operatorname{Var}(X) = (1-p)/p^2$.

B3 Let $X \sim \operatorname{Po}(\lambda)$.

(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 $\operatorname{Var}(X) = \lambda$. 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 Lecture 6, 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 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.

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 28~~ Tuesday 29 November** ~~at the beginning~~ of Week 9. Submission will be via Gradescope. Your work will be marked by your tutor and returned on ~~Monday 5~~ Wednesday 7 December, 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.

Because marking time will be reduced due to the UCU strike on Wednesday 30 November, Questions C1(d) and C2(c) are no longer required to be submitted. You may choose to complete for your own practice.

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 expected 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 $\operatorname{Cov}(aX, Y) = a\operatorname{Cov}(X, Y)$. You may find it helpful to remember that if $\mathbb{E}X = \mu_X$, then $\mathbb{E}aX = a\mu_X$.

(b) Show that $\operatorname{Cov}(X + b, Y) = \operatorname{Cov}(X, Y)$.

~~Now let X, Y, Z be *independent* random variables with common variance σ^2.~~

~~**(c)** Find the value of 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 **A4.** (a) $\frac{1}{15}$ (d) No

Chapter 15

Continuous random variables

15.1 What is a continuous random variable?

In the previous six lectures, we have looked at discrete random variables, whose range is a finite or countably infinite set of separate discrete values. Discrete random variables can be used as a model for “count data”.

In this section and the next, we will instead look at continuous random variables, whose range is an uncountable set, a continuum of gradually varying values. Continuous random variables can be used as a model for “measurement data”. For example:

- The assets of a bank at the end of this year could be modelled as a continuous random variable with range the real numbers $\mathbb{R}$, where positive numbers represent credit and negative numbers represent debt.
- The amount of time a machine in a factory works for before breaking down could be modelled as a continuous random variable with range the positive real numbers $\mathbb{R}_+ = \{x \in \mathbb{R} : x \geq 0\}$.
- The unemployment rate in the UK next January, as a proportion of the population, could be measured as a continuous random variable with range the interval $[0, 1] = \{x \in \mathbb{R} : 0 \leq x \leq 1\}$.

Imagine firing an arrow at a large target. We could ask “What’s the probability that the arrow exactly hits some point?” – but this question is difficult to answer. What do we mean by a point? If we mean a mathematically-idealised infinitesimally small point, then I think we’d have to say that the probability is 0. What makes more sense is too take a section of the target – perhaps a small circle in the middle, called the “bulls-eye” – and ask what is the probability that the arrow lands in the area of the bulls-eye. Then we could (at least in theory) answer that question – a good archer would have quite a high probability of landing the arrow in the bulls-eye, while a poor archer would have a smaller chance.

Similarly, imagine picking a random real number between 0 and 1. We could ask "What is the probability that the random number is *exactly* $1/\sqrt{2} = 0.7071068...$?" But that probability, if it means anything, must be 0. It makes more sense to take an interval of numbers – say, $[0.7, 0.8]$, the interval from 0.7 to 0.8 – and ask what the probability is of the random number being in that interval.

This is how continuous random variables work. The probability a continuous random variable X *exactly* hits some value x is $\mathbb{P}(X = x) = 0$. But we *can* find the probability $\mathbb{P}(a \leq X \leq b)$ that X lies in a certain interval and work with that.

15.2 Probability density functions

With a continuous random variable, the probability of *exactly* getting any particular outcome $X = x$ is 0. However, we can express the "intensity" of probability *around* x by $f_X(x)$, where f_X is called the "probability density function". The implied metaphor here is that for discrete random variables, we have probability "mass" *at* the point x, whereas for continuous random variables, we have a "density" of probability *around* x.

Definition 15.1. A random variable X is called a **continuous random variable** if the probability of landing in any interval between a and b, for $a \leq b$, can be written as

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)\,\mathrm{d}x,$$

for some non-negative function f_X. The function f_X is called the **probability density function** (or **PDF**).

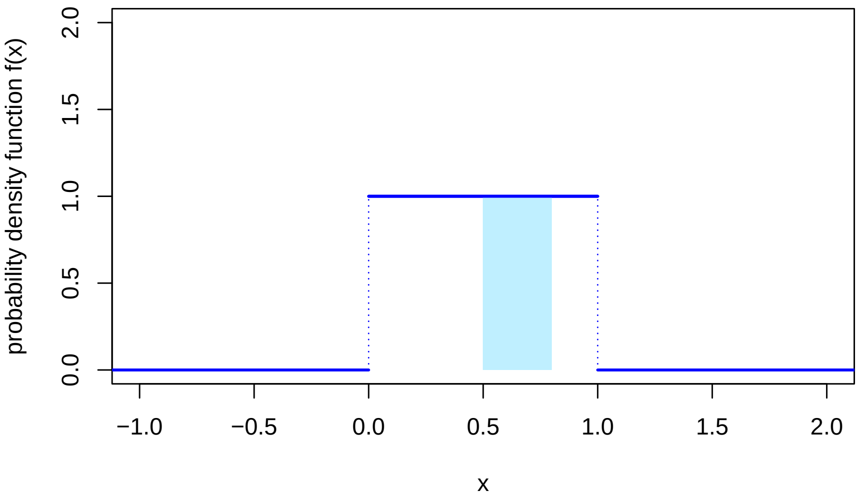
In other words, the probability that X is between a and b is the area under the curve of the PDF $f_X(x)$ between $x = a$ and $x = b$.

As with PMFs, when it's obvious what random variable we're dealing with, we omit the subscript X on the PDF f_X.

Example 15.1. Let X be a continuous random variable with PDF

$$f(x) = 1 \qquad \text{for } 0 \leq x \leq 1$$

and $f(x) = 0$ otherwise. This represents a random number between 0 and 1, where the intensity of the probability is equal across the whole interval. This is known as a **continuous uniform distribution**.



What is the probability that X is between 0.5 and 0.8?

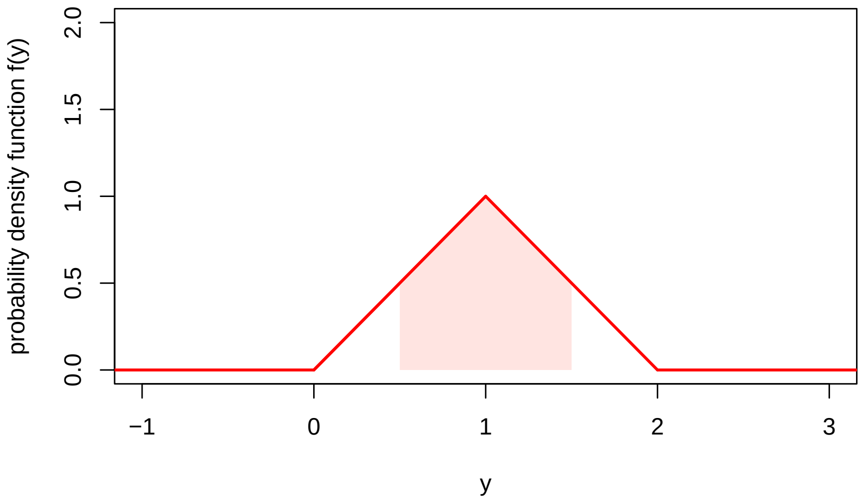
We can calculate this using the definition above. We have

$$\begin{aligned}\mathbb{P}(0.5 \leq X \leq 0.8) &= \int_{0.5}^{0.8} f(x)\,\mathrm{d}x\\ &= \int_{0.5}^{0.8} 1\,\mathrm{d}x\\ &= [x]_{0.5}^{0.8}\\ &= 0.8 - 0.5\\ &= 0.3.\end{aligned}$$

Example 15.2. Let Y be a continuous random variable with PDF

$$f(y) = \begin{cases} y & \text{for } 0 \leq y \leq 1\\ 2-y & \text{for } 1 < y \leq 2 \end{cases}$$

and $f(y) = 0$ otherwise. This represents a continuous value between 0 and 2 where the probability intensity is highest in the middle around 1 and is lower at the edges near 0 and 2.



What is the probability X is between $\frac12$ and $\frac32$?

As before, we have

$$\mathbb{P}(\tfrac12 \leq Y \leq \tfrac32) = \int_{\frac12}^{\frac32} f(y)\,\mathrm{d}y.$$

But this time we have to be careful, because $f(y)$ has different expressions below 1 and above 1. We will split the integral up into two parts based on this, to get

$$\begin{aligned}\mathbb{P}(\tfrac12 \leq Y \leq \tfrac32) &= \int_{\frac12}^{1} f(y)\,\mathrm{d}y + \int_{1}^{\frac32} f(y)\,\mathrm{d}y \\ &= \int_{\frac12}^{1} y\,\mathrm{d}y + \int_{1}^{\frac32} (2-y)\,\mathrm{d}y \\ &= \left[\tfrac12 y^2\right]_{\frac12}^{1} + \left[2y - \tfrac12 y^2\right]_{1}^{\frac32} \\ &= \tfrac12 - \tfrac18 + (\tfrac62 - \tfrac98) - (2 - \tfrac12) \\ &= \tfrac34.\end{aligned}$$

15.3 Properties of continuous random variables

The good news is that almost all of the properties we know and love about discrete distributions also follow through for continuous distribution – except you swap the PMF for the PDF and swap sums for integrals.

Discrete random variables	Continuous random variables
A discrete random variable X is defined by a **probability mass function** (PMF) $p(x)$, which represents the probability of getting exactly x.	A continuous random variable X is defined by a **probability density function** (PDF) $f(x)$, which represents the intensity of probability around x.
The PMF is positive, in that $p(x) \geq 0$ for all x.	The PDF is positive, in that $f(x) \geq 0$ for all x.
The PMF sums to 1, in that $\sum_x p(x) = 1$.	The PDF integrates to 1 in that $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$.
The **cumulative distribution function** (CDF) is $F(x) = \mathbb{P}(X \leq x)$, and is given by a sum $F(x) = \sum_{y \leq x} p(y)$.	The **cumulative distribution function** (CDF) is $F(x) = \mathbb{P}(X \leq x)$, and is given by an integral $F(x) = \int_{-\infty}^{x} f(y)\,\mathrm{d}y$.
The **expectation** is the sum $\mathbb{E}X = \sum_x x\,p(x)$.	The **expectation** is the integral $\mathbb{E}X = \int_{-\infty}^{\infty} x\,f(x)\,\mathrm{d}x$.
The expectation of a function $g(X)$ of X is the sum $\mathbb{E}g(X) = \sum_x g(x)\,p(x)$.	The expectation of a function $g(X)$ of X is the integral $\mathbb{E}g(X) = \int_{-\infty}^{\infty} g(x)\,f(x)\,\mathrm{d}x$.
Linearity of expectation says that $\mathbb{E}(aX+b) = a\mathbb{E}X + b$.	Linearity of expectation says that $\mathbb{E}(aX+b) = a\mathbb{E}X + b$.
The **variance** is $\mathrm{Var}(X) = \mathbb{E}(X-\mu)^2$, which also has the computational formula $\mathrm{Var}(X) = \mathbb{E}X^2 - \mu^2$.	The **variance** is $\mathrm{Var}(X) = \mathbb{E}(X-\mu)^2$, which also has the computational formula $\mathrm{Var}(X) = \mathbb{E}X^2 - \mu^2$.

Note, however, one property that doesn't follow through: Because, for a PMF, $p(x) = \mathbb{P}(X = x)$ represented a probability, we had $p(x) \leq 1$ for all x. However, because, for a PDF, $f(x)$ only represents intensity of probability, there's no contradiction to having $f(x) > 1$ (although keeping the integral to 1 means that we can't have $f(x) > 1$ too much). So $f(x) = 10$ for $0 < x < 0.1$ and $f(x) = 0$ otherwise is a perfectly legitimate PDF, for example.

Example 15.3. Let's return to the case where X be a continuous uniform distribution, with

$$f(x) = 1 \qquad \text{for } 0 \leq x \leq 1$$

and $f(x) = 0$ otherwise. Let's go through the properties from the table above.

First, it's clear that $f(x) \geq 0$ for all x.

Second, the PDF does indeed integrate to 1, because

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = \int_{0}^{1} 1\,\mathrm{d}x = [x]_0^1 = 1.$$

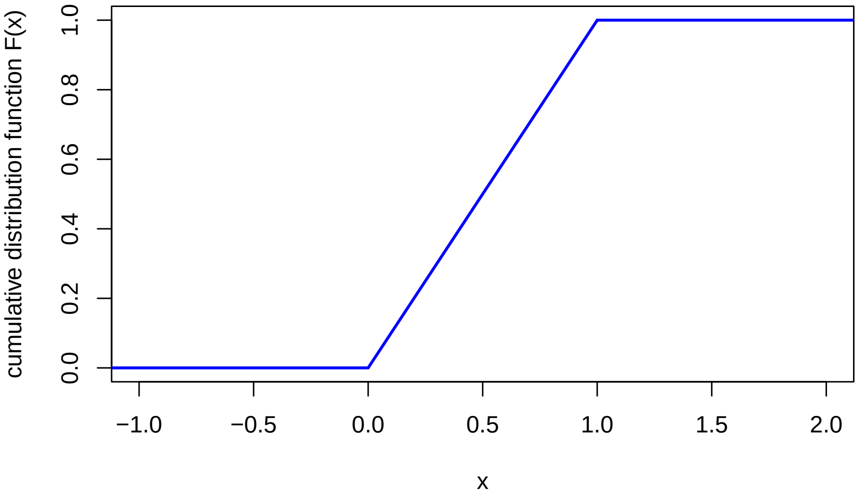
Because this PDF is zero below 0 and above 1, we only had to integrate between 0 and 1, with the rest of the integral over the real line being 0.

Third, the CDF F. It's clear that $F(x) = \mathbb{P}(X \leq x) = 0$ for $x < 0$, and $F(x) = \mathbb{P}(X \leq x) = 1$ for $x > 1$. In between, we have

$$F(x) = \int_{-\infty}^{x} f(y)\,\mathrm{d}y = \int_{0}^{x} 1\,\mathrm{d}y = [y]_0^x = x.$$

So, altogether, the CDF is

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



Fourth, the expectation is

$$\mathbb{E}X = \int_{\infty}^{\infty} x\,f(x)\,\mathrm{d}x = \int_{0}^{1} x\,\mathrm{d}x = \left[\tfrac{1}{2}x^2\right]_0^1 = \tfrac{1}{2} - 0 = \tfrac{1}{2}.$$

Finally, to calculate the variance using the computational formula $\mathrm{Var}(X) = \mathbb{E}X^2 - \mu^2$, we first need $\mathbb{E}X^2$. This is

$$\mathbb{E}X^2 = \int_{\infty}^{\infty} x^2\,f(x)\,\mathrm{d}x = \int_{0}^{1} x^2\,\mathrm{d}x = \left[\tfrac{1}{3}x^3\right]_0^1 = \tfrac{1}{3} - 0 = \tfrac{1}{3}.$$

So, the variance is

$$\mathrm{Var}(X) = \mathbb{E}X^2 - \mu^2 = \tfrac{1}{3} - \left(\tfrac{1}{2}\right)^2 = \tfrac{1}{3} - \tfrac{1}{4} = \tfrac{1}{12}.$$

Example 15.4. Let's also return to the "triangular" PDF from Example 15.2,

$$f(y) = \begin{cases} y & \text{for } 0 \leq y \leq 1 \\ 2 - y & \text{for } 1 < y \leq 2 \end{cases}$$

and $f(y) = 0$ otherwise. We'll just do the CDF and the expectation. (You can do the others yourself, if you like.)

For the CDF, it's clear that $F(y) = 0$ for $y < 0$ and $F(y) = 1$ for $y > 2$. Again, we split the $0 \leq y \leq 1$ case and the $1 < y \leq 2$ case. In the first case, for $0 \leq y \leq 1$, we have

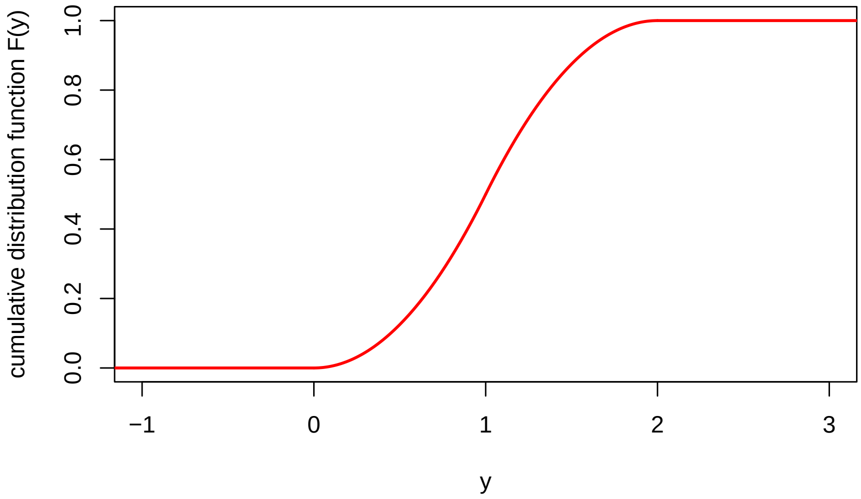
$$\begin{aligned} F(x) &= \int_{-\infty}^{y} f(z)\,\mathrm{d}z \\ &= \int_{0}^{y} z\,\mathrm{d}z \\ &= \left[\tfrac{1}{2}z^2\right]_0^y \\ &= \tfrac{1}{2}y^2. \end{aligned}$$

In the second case, for $1 < y \leq 2$, we have

$$\begin{aligned} F(x) &= \int_{-\infty}^{y} f(z)\,\mathrm{d}z \\ &= \int_{0}^{1} z\,\mathrm{d}z + \int_{1}^{y} (2-z)\,\mathrm{d}z \\ &= \left[\tfrac{1}{2}z^2\right]_0^1 + \left[2z - \tfrac{1}{2}z^2\right]_1^y \\ &= \tfrac{1}{2} - 0 + 2y - \tfrac{1}{2}y^2 - 2 + \tfrac{1}{2} \\ &= 2y - \tfrac{1}{2}y^2 - 1. \end{aligned}$$

Hence, the CDF is

$$F(y) = \begin{cases} 0 & \text{for } y < 0 \\ \tfrac{1}{2}y^2 & \text{for } 0 \leq y \leq 1 \\ 2y - \tfrac{1}{2}y^2 - 1 & \text{for } 1 < y \leq 2 \\ 1 & \text{for } y > 2. \end{cases}$$



For the expectation, we have

$$\begin{aligned}\mathbb{E}Y &= \int_{-\infty}^{\infty} y\, f(y)\, \mathrm{d}y \\ &= \int_0^1 y^2 \mathrm{d}y + \int_1^2 y(2-y)\, \mathrm{d}y \\ &= \left[\tfrac13 y^3\right]_0^1 + \left[y^2 - \tfrac13 y^3\right]_1^2 \\ &= \tfrac13 - 0 + 4 - \tfrac83 - 1 + \tfrac13 \\ &= 1.\end{aligned}$$

Summary

- A continuous random variable is defined by its probability density function f, where
$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)\, \mathrm{d}x.$$
- Most properties of discrete random variables hold, with the PMF replaced by the PDF, and sums by integrals.
- For example, the expectation is $\mathbb{E}X = \displaystyle\int_{-\infty}^{\infty} x\, f(x)\, \mathrm{d}x$.

Chapter 16

Normal distribution

When we studied discrete random variables, we studied a number of important families of distributions: the Bernoulli, binomial, geometric, and Poisson distributions. In this lecture, we'll look at an extremely important distribution: the "normal" (or "Gaussian") distribution.

16.1 Definition of the normal distribution

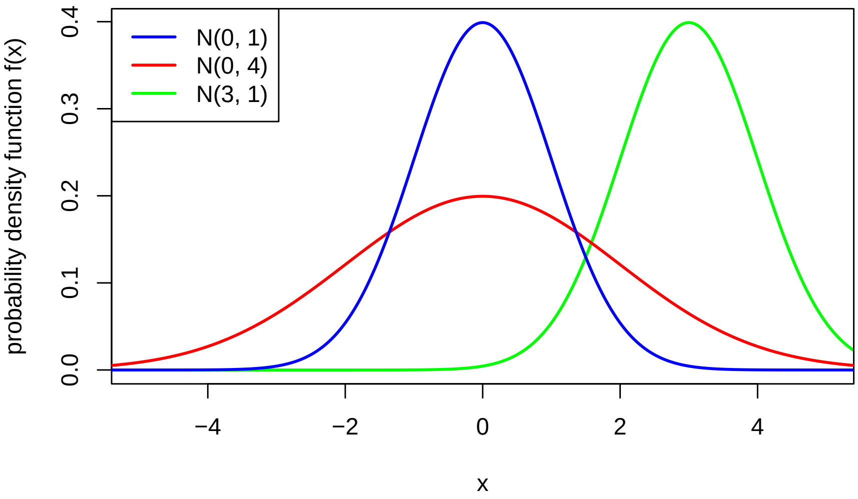
Definition 16.1. If X is a continuous random variable with PDF

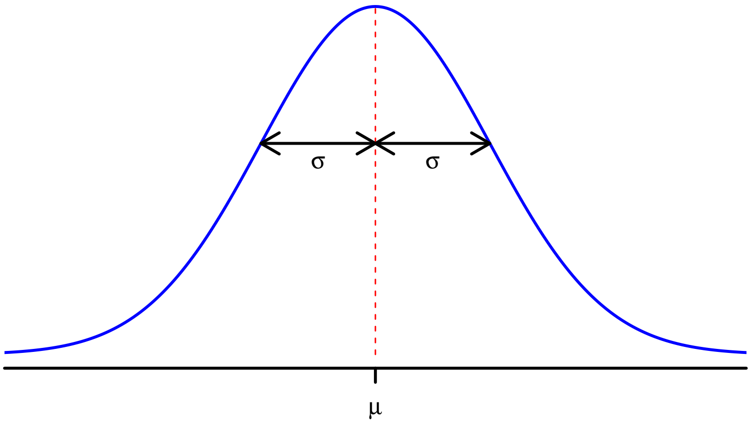
$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

then we say that X has the **normal distribution** with expectation μ and variance $\sigma^2 > 0$, and write $X \sim \mathrm{N}(\mu, \sigma^2)$.

(Many people call μ the "mean", which is a slight misnomer.)

This PDF is the famous "bell curve", where the centre of the bell is at $x = \mu$ and the width of the bell is controlled by the value of σ^2. Note also that the PDF is symmetric about μ.





One important special case is $\mu = 0$ and $\sigma^2 = 1$, in which case we say that $Z \sim \mathrm{N}(0, 1)$ has the **standard normal distribution**. We typically write ϕ (lower-case "phi"), where

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-z^2/2}$$

for the PDF of a standard normal distribution, and write Φ (upper-case "Phi"), where

$$\Phi(z) = \mathbb{P}(Z \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \mathrm{e}^{-y^2/2} \, \mathrm{d}y$$

for the CDF of a standard normal distribution.

The normal distribution is a very widely used distribution for modelling many things in real life.

- Measurement error with scientific instruments is typically modelled as a normal distribution with expectation $\mu = 0$. The more precise the instrument, the lower the value of the variance σ^2.
- According to a poll a few years ago, the height of MATH1712 students in centimetres can be modelled well by a normal distribution with expectation $\mu = 172$ and variance $\sigma^2 = 86$.
- In financial models, it is often assumed that the logarithm of the daily change in a stock price follows a normal distribution. In this context, the expectation μ is known as the "drift" and the standard deviation σ as the "volatility". This "log-normal" model is the basis of the famous Black–Scholes model of financial markets.

More generally, and for reasons we will come back to later, the normal distribution is good for modelling things where lots of little effects add together to make a bigger effect. We will also see later that many other distributions can be approximated by a normal distribution.

It's generally difficult, or even impossible, to directly calculate probabilities of events concerning the normal distribution. Instead, one must use numerical approximations. We will discuss these further later in this section.

16.2 Properties of the normal distribution

Theorem 16.1. *Let $X \sim \mathrm{N}(\mu, \sigma^2)$ be a normally distributed random variable. Then:*

1. *$f_X(x)$ is indeed a PDF, in that $\displaystyle\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = 1$;*
2. *$\mathbb{E}X = \mu$;*
3. *$\mathrm{Var}(X) = \sigma^2$.*

In particular, if $Z \sim \mathrm{N}(0,1)$ is a standard normal distribution, then $\mathbb{E}Z = 0$ and $\mathrm{Var}(Z) = 1$.

We'll give (non-examinable) proofs of these soon. But first we'll note one other thing.

Let $X \sim \mathrm{N}(\mu, \sigma^2)$, and consider the random variable $Y = aX + b$. Then we know that

$$\begin{aligned}\mathbb{E}(aX + b) &= a\mu + b,\\ \mathrm{Var}(aX + b) &= a^2\sigma^2.\end{aligned}$$

In fact, it can be shown that $aX+b$ is normally distributed too; that is, $aX+b \sim \mathrm{N}(a\mu+b, a^2\sigma^2)$. Importantly, if we take $a = 1/\sigma$ and $b = -\mu/\sigma$, then we see that

$$Z = \frac{X-\mu}{\sigma} \sim \mathrm{N}(0,1).$$

In other words, we can stretch and scale any normal random variable to turn it into a standard normal random variable. This is known as "standardisation" and will be useful later.

We can also use standardisation to help us prove Theorem 16.1.

Proof. (Non-examinable) By using standardisation, it suffices to prove the theorem for a standard normal random variable $X \sim \mathrm{N}(0,1)$.

For part 1, we need to show that

$$I = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{e}^{-x^2/2}\,\mathrm{d}x = 1.$$

To prove this we use one of the most outrageous tricks in mathematics! The first part of the trick is that, instead of calculating the integral itself I, we can instead calculate the square of the integral I^2, which we also need to show is equal to 1. This is

$$\begin{aligned} I^2 &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{e}^{-x^2/2}\,\mathrm{d}x \times \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{e}^{-y^2/2}\,\mathrm{d}y \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathrm{e}^{-x^2/2}\,\mathrm{e}^{-y^2/2}\,\mathrm{d}x\,\mathrm{d}y \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathrm{e}^{-(x^2+y^2)/2}\,\mathrm{d}x\,\mathrm{d}y. \end{aligned}$$

The second part of the outrageous trick is notice that the appearance of x^2+y^2 suggests it might be useful to transfer from cartesian coordinates (x,y) to polar coordinates (r,θ). Recalling that $x^2+y^2=r^2$ and $\mathrm{d}x\,\mathrm{d}y = r\,\mathrm{d}r\,\mathrm{d}\theta$, we have

$$\begin{aligned} I^2 &= \frac{1}{2\pi} \int_0^{2\pi} \int_0^{\infty} \mathrm{e}^{-r^2/2}\,r\,\mathrm{d}r\,\mathrm{d}\theta \\ &= \frac{1}{2\pi}\,2\pi \int_0^{\infty} r\,\mathrm{e}^{-r^2/2}\,\mathrm{d}r \\ &= \left[-\mathrm{e}^{-r^2/2}\right]_0^{\infty} \\ &= -0-(-1) \\ &= 1, \end{aligned}$$

and we're done.

For part 2, we need to show that $\mathbb{E}X = 0$. We have

$$\begin{aligned}\mathbb{E}X &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x\,\mathrm{e}^{-x^2/2}\,\mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}} \left[-\mathrm{e}^{-x^2/2}\right]_{-\infty}^{\infty} \\ &= -0 - (-0) \\ &= 0,\end{aligned}$$

as required.

For part 3, we need to show that $\mathbb{E}X^2 = 1$. Using integration by parts with $u = x$, $v' = x\,\mathrm{e}^{-x^2/2}$, we have

$$\begin{aligned}\mathbb{E}X^2 &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2\,\mathrm{e}^{-x^2/2}\,\mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}} \left[-x\mathrm{e}^{-x^2/2}\right]_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{e}^{-x^2/2}\,\mathrm{d}x \\ &= 0 + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{e}^{-x^2/2}\,\mathrm{d}x.\end{aligned}$$

But this integral on the right is just the integral I of the PDF as above, which we know equals 1, as required. □

There is one last property of the normal distribution that we won't use directly in this module, but is perhaps worth knowing anyway.

Theorem 16.2. *If $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$ are independent, then*

$$X + Y \sim \mathrm{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2).$$

16.3 Calculations using R

We will try to answer a number of questions about the normal distribution.

Question 1. *A fiberoptic fibre is manufactured with an average width of 8 nanometres (nm), with a standard deviation of 0.04 nm. Fibres that are wider than 8.1 nm fail testing and must be discarded. If the manufactured width is modelled as normally distributed, then what proportion of fibres pass the test?*

Let $X \sim \mathrm{N}(8, 0.04^2)$ denote the width of a random fibre, measured in nanometres. Then this question required us to find

$$F(8.15) = \mathbb{P}(X \leq 8.1) = \frac{1}{\sqrt{2\pi \times 0.04^2}} \int_{-\infty}^{8.1} \exp\left(-\frac{(x-8)^2}{2 \times 0.04^2}\right) \mathrm{d}x.$$

Unfortunately, it is not possible to calculate this integral exactly. However, computers can approximate this integral very accurately and very quickly. In R, this is done with the `pnorm()` function, which calculates the CDF of a normal distribution. `pnorm()` typically takes three arguments:

1. the first argument is the value x at which we wish to evaluate the CDF;
2. the second argument is the expectation μ of the normal distribution;
3. the third argument is the standard deviation σ of the normal distribution. (Note that this third argument is the *standard deviation* σ and not the variance σ^2. This is an easy mistake to make!)

So here, the number we want is

```
pnorm(8.1, 8, 0.04)
```

```
## [1] 0.9937903
```

We see that roughly 99.4% of fibres pass the test.

Question 2. *Let* $Z \sim \mathrm{N}(0,1)$*. What is* $\mathbb{P}(Z \leq 1.45)$*?*

This is asking for $\Phi(1.45) = \mathbb{P}(Z \leq 1.45)$. This is:

```
pnorm(1.45, 0, 1)
```

```
## [1] 0.9264707
```

But in fact, the standard normal distribution CDF Φ is so common that R allows you to omit the values of μ and σ if they are 0 and 1 respectively. So you can save yourself a few keystrokes by simply writing:

```
pnorm(1.45)
```

```
## [1] 0.9264707
```

Question 3. *Let* $Z \sim \mathrm{N}(0,1)$*. What is* $\mathbb{P}(Z > 0.33)$*?*

This is asking for the upper-tail probability. The direct way to get R to solve this is to use the `lower.tail = FALSE` option that we discussed in R Worksheet 7. That is, we use:

```
pnorm(0.33, lower.tail = FALSE)
```

```
## [1] 0.3707
```

Alternatively, we could use the fact that $\mathbb{P}(Z > z) = 1 - \mathbb{P}(Z \leq z) = 1 - \Phi(z)$. Then we could equally well calculate this as

```
1 - pnorm(0.33)
```

```
## [1] 0.3707
```

Question 4. *We return to the fiberoptic model $X \sim \mathrm{N}(8, 0.04^2)$ from Question 1. Fibres can be awarded a special "high quality" stamp if their width is between 7.95 and 8.05 nm. What proportion of these fibres qualify?*

This is asking for $\mathbb{P}(7.95 \leq X \leq 8.05)$. But we can calculate this as

$$\mathbb{P}(7.95 \leq X \leq 8.05) = \mathbb{P}(X \leq 8.05) - \mathbb{P}(X < 7.95) = F(8.05) - F(7.95).$$

(Formally, this is because

$$\{X < 7.95\} \cup \{7.95 \leq X \leq 8.05\} = \{X \leq 8.05\}$$

is a disjoint union, so we can use Axiom 3.)

So the proportion of qualifying fibres is

```r
mu <- 8
sigma <- 0.04
pnorm(8.05, mu, sigma) - pnorm(7.95, mu, sigma)
```

```
## [1] 0.7887005
```

or about 79%.

Question 5. *We stay with the fiberoptic model $X \sim \mathrm{N}(8, 0.04^2)$ from Questions 1 and 4. The manufacturer wants to be able to advertise that 99.9% of their fibres are between lower and upper limits x and y. What values of x and y can they promise?*

Is F is the CDF of this distribution, then we are looking for x and y such that $F(x) = 0.0005$ and $F(y) = 0.9995$. That way, $F(y) - F(x) = 0.999$, so we have 99.9% of fibres within that interval and 0.05% outside either side.

You may remember from R Worksheet 7 that the inverse F^{-1} of the CDF is called the **quantile function**. Here, we want $F^{-1}(0.0005)$ and $F^{-1}(0.9995)$. The quantile function for the normal distribution in R is `qnorm()`. (It also has a `lower.tail = FALSE` option, which is sometimes useful.) So we can use

```r
mu <- 8
sigma <- 0.04
c(qnorm(0.0005, mu, sigma), qnorm(0.9995, mu, sigma))
```

```
## [1] 7.868379 8.131621
```

We see that we can guarantee that 99.9% of fibres are between roughly 7.87 and 8.13 nm wide.

16.4 Calculations using statistical tables

Doing normal calculations with R is all very well. But what if you accidentally built a time machine and got transported back to Victorian times. Then how would you perform calculations with the normal distribution?

In the olden days, someone would (using some enormous computer the size of a room, or whatever) calculate lots of values of $\Phi(x)$, the CDF of the standard normal distribution, and publish them in a book of statistical tables. An example of this is **this page of normal distribution tables** [PDF] that will appear on the final page of your exam. (Like the Victorian times, your exam is another place R will not be available but statistical tables will be.)

We will return to the same questions we answered in the previous subsection, although in a slightly different order.

Question 2. *Let $Z \sim \mathrm{N}(0,1)$. What is $\mathbb{P}(Z \leq 1.45)$?*

As we noted before, this is asking for $\Phi(1.45) = \mathbb{P}(Z \leq 1.45)$. Consulting the statistical tables, we see that the value of $\Phi(1.45)$ is listed on the table. Specifically, we see from column 3, row 10 of Table 1 that $\Phi(1.45) = 0.9265$. This is the same value as we got from R (although we get fewer decimal places from the table).

Question 1. *A fiberoptic fibre is manufactured with an average width of 8 nanometres (nm), with a standard deviation of 0.04 nm. Fibres that are wider than 8.1 nm fail testing and must be discarded. If the manufactured width is modelled as normally distributed, then what proportion of fibres pass the test?*

If $X \sim \mathrm{N}(8, 0.04^2)$, then this asks for $F_X(8.1) = \mathbb{P}(X \leq 8.1)$. However, unfortunately the statistical tables only have the CDF Φ for the standard normal distribution $\mathrm{N}(0,1)$. So we are going to have "standardise" X; that is, convert X to a standard normal distribution. Recall from above that we standardise a normal random variable by subtracting the expectation μ and dividing by the standard deviation σ. So in this case, we have

$$Z = \frac{X-\mu}{\sigma} = \frac{X-8}{0.04} \sim \mathrm{N}(0,1).$$

Using this, we can write

$$\mathbb{P}(X \leq 8.1) = \mathbb{P}\left(\frac{X-8}{0.04} \leq \frac{8.1-8}{0.04}\right) = \mathbb{P}(Z \leq 2.5) = \Phi(2.5).$$

We can then look up $\Phi(2.5)$ in Table 1. We see from the first row of the last column that $\Phi(2.5) = 0.9938$. This matches the answer we got from R.

Question 3. *Let $Z \sim \mathrm{N}(0,1)$. What is $\mathbb{P}(Z > 0.33)$?*

The statistical tables only have $\Phi(z) = \mathbb{P}(Z \leq z)$. But as we noted above, $\mathbb{P}(Z > 0.33) = 1 - \Phi(0.33)$. The tables don't have $\Phi(0.33)$ either, though, because they jump straight from $\Phi(0.30)$ to $\Phi(0.35)$. We have two choices of what to do here.

First choice, which is appropriate when an approximate answer will suffice, is simply to take the nearest value in the table, which here is 0.35. Hence

$$\mathbb{P}(Z > 0.33) = 1 - \Phi(0.33) \approx 1 - \Phi(0.35) = 1 - 0.6368 = 0.3632.$$

This is pretty close to the true answer 0.3707 we saw before: about a 2% error.

Second choice, which is more work but gets a more accurate answer, is to use interpolation. We know from the table that $\Phi(0.30) = 0.6179$ and $\Phi(0.35) = 0.6368$. To "interpolate", we assume that the graph of Φ follows a straight line between $(0.30, 0.6179)$ and $(0.35, 0.6368)$. (In fact, Φ has a slightly curve, so isn't *quite* straight.) As the statistical tables state, the interpolation is to take

$$\Phi(x) = \frac{x_2 - x}{x_2 - x_1}\Phi(x_1) + \frac{x - x_1}{x_2 - x_1}\Phi(x_2).$$

In our case, if we take $x_1 = 0.30$ and $x_2 = 0.35$ as the interpolation points for $x = 0.33$, we get the approximation

$$\Phi(0.33) = 0.4\Phi(0.30) + 0.6\Phi(0.35) = 0.4 \times 0.6179 + 0.6 \times 0.6368 = 0.6292$$

This is off by only 0.01%; a very accurate approximation.

On problem sheets or in the exam, you will be told if an interpolation is necessary.

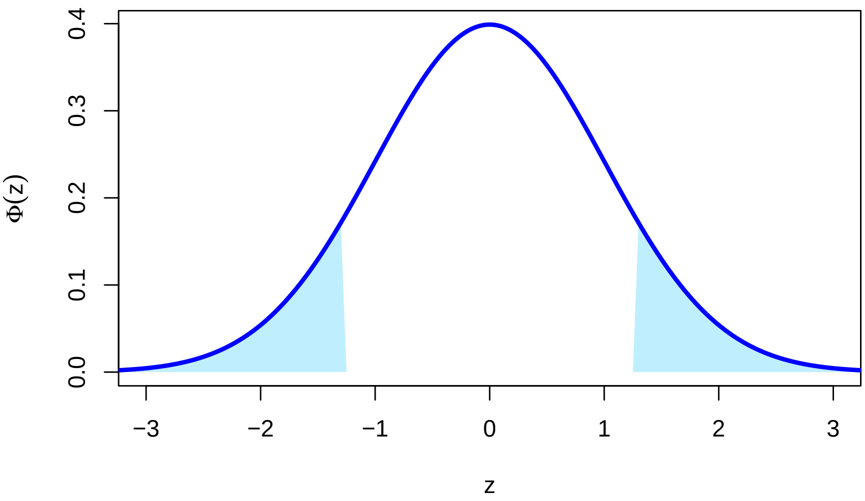
Question 4. *We return to the fiberoptic model $X \sim \mathrm{N}(8, 0.04^2)$ from Question 1. Fibres can be awarded a special "high quality" stamp if their width is between 7.95 and 8.05 nm. What proportion of these fibres qualify?*

As noted above, this is asking for $\mathbb{P}(7.95 \leq X \leq 8.05)$. To allow us to use our statistical tables, we will have to standardise. We get

$$\begin{aligned}\mathbb{P}(7.95 \leq X \leq 8.05) &= \mathbb{P}\left(\frac{7.95-8}{0.04} \leq \frac{X-8}{0.04} \leq \frac{8.05-8}{0.04}\right)\\ &= \mathbb{P}(-1.25 \leq Z \leq 1.25)\\ &= \Phi(1.25) - \Phi(-1.25).\end{aligned}$$

We can find $\Phi(1.25) = 0.8944$ from the table. But the table only gives $\Phi(x)$ for positive x, so we can't look up $\Phi(-1.25)$.

Instead, we can use the symmetry of the normal distribution. Because the standard normal is symmetric about 0, we have that $\mathbb{P}(Z \leq -1.25) = \mathbb{P}(Z > 1.25)$.



Therefore, we have

$$\Phi(-1.25) = \mathbb{P}(Z > 1.25) = 1 - \Phi(1.25) = 1 - 0.8944 = 0.1056$$

Putting this all together, we get

$$\mathbb{P}(7.95 \leq X \leq 8.05) = 0.8944 - 0.1056 = 0.7888,$$

which is the same thing as we got from R (up to a small rounding error in the fourth decimal place).

Question 5. *We stay with the fiberoptic model $X \sim \mathrm{N}(8, 0.04^2)$ from Questions 1 and 4. The manufacturer wants to be able to advertise that 99.9% of their fibres are between lower and upper limits x and y. What values of x and y can they promise?*

Recall that this meant we were looking for the quantiles $F^{-1}(0.0005)$ and $F^{-1}(0.9995)$; that is, the values x and y such that $\mathbb{P}(X \leq x) = 0.0005$ and $\mathbb{P}(X \leq x) = 0.9995$. Table 2 of the statistical tables does show us some quantiles for a standard normal. How can we use these?

Let's start with the second case. The key here is to "undo" the standardisation. That is, if $Z \sim \mathrm{N}(0, 1)$, then $X = \sigma Z + \mu \sim \mathrm{N}(\mu, \sigma^2)$. The table tells us that $\Phi^{-1}(0.9995) = 3.2905$; that is, that $\mathbb{P}(Z \leq 3.2905) = 0.9995$. Then by "un-standardising", we have

$$0.9995 = \mathbb{P}(Z \leq 3.2905) = \mathbb{P}(0.04Z + 8 \leq 0.04 \times 3.2905 + 8) = \mathbb{P}(X \leq 8.1316).$$

This the upper quantile we are after is 8.1316.

For the lower quantile, we can use symmetry again. Thus the $0.0005 = 1-0.9995$ quantile for Z is minus the previous quantile; that is, -3.2905. Hence the lower quantile we want is

$$0.04 \times (-3.2905) + 8 = 7.8684.$$

These match the answers we got with R.

I feel I shouldn't finish with this subsection before addressing the following question some readers may be asking themselves: *Now that we have R (and other computing methods), what's the point learning to answer questions using statistical tables?* I might suggest a few possible answers to this question:

1. Although using statistical tables is an archaic skill, in order to use the statistical tables, you will need to know and be able to apply many facts about probability distributions in general and the normal distribution in particular. So this is a good way to learn those facts and practice their application.
2. Someone has to write the computer program, and these people need to be able to do the sorts of conversions we will learn about here. So these are useful skills for mathematician–programmers to learn.
3. Being able to standardise normal distributions, approximate other distributions by normal distributions (see Lecture 18), and so on, are actually important to be able to solve purely mathematical problems, quite outside of merely performing calculations.
4. Yes, you are right, this is a pointless skill for us to teach you.

I am mostly convinced by answers 1 to 3, although I must admit that answer 4 isn't totally without merit.

Summary

- The normal distribution has PDF
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$
It has expectation μ and variance σ^2.
- The standard normal distribution has $\mu = 0$ and $\sigma^2 = 1$.
- The CDF of a normal distribution can be calculated in R with the `pnorm()` function. It can also be calculated using statistical tables and by "standardising".

Chapter 17

Exponential distribution and multiple continuous random variables

This lecture is really two mini-lectures stuck together. In the first mini-lecture, we look at another important continuous distribution, the exponential distribution. In the second mini-lecture, we look at the theory of multiple continuous random variables, and find it's very similar to what we already know about multiple discrete random variables.

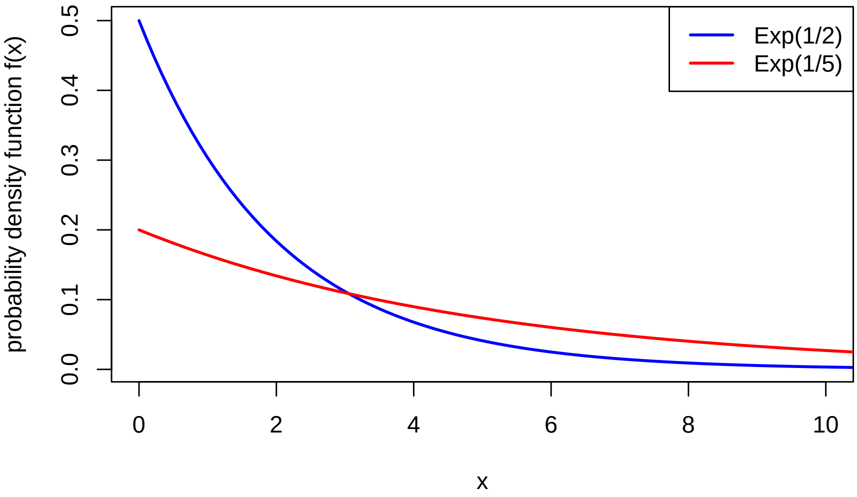
17.1 Exponential distribution

An important continuous distribution is the exponential distribution. The exponential distribution is often used to represent lengths of time: for example, the time between radioactive particles decaying, the time between eruptions of a volcano, or the time between buses arriving at a bus stop.

Definition 17.1. A continuous random variable X is said to have the **exponential distribution with rate** $\lambda > 0$ if it has the PDF

$$f(x) = \lambda \mathrm{e}^{-\lambda x} \qquad \text{for } x \geq 0,$$

and 0 otherwise. We write $X \sim \mathrm{Exp}(\lambda)$.



Example 17.1. *The length of time in years that a lightbulb works before needing to be replaced is modelled as an exponential distribution with rate $\lambda = 2$. What is the probability the lightbulb needs replacing within a year?*

If $X \sim \mathrm{Exp}(2)$ is the lifetime of the lightbulb, we seek $\mathbb{P}(X \leq 1)$. This is

$$\int_{-\infty}^{1} f(x)\,\mathrm{d}x = \int_{0}^{1} 2\mathrm{e}^{-2x}\,\mathrm{d}x = \big[-\mathrm{e}^{-2x}\big]_0^1 = -\mathrm{e}^{-2} - (-1) = 1 - \mathrm{e}^{-2} = 0.864.$$

Theorem 17.1. *Suppose $X \sim \mathit{Exp}(\lambda)$. Then:*

1. *f is indeed a PDF, in that* $\displaystyle\int_0^\infty f(x)\,\mathrm{d}x = 1$*;*
2. *the CDF of X is $F(x) = 1 - \mathrm{e}^{-\lambda x}$;*
3. *the expectation of X is* $\mathbb{E}X = \dfrac{1}{\lambda}$*;*
4. *the variance of X is* $\mathrm{Var}(X) = \dfrac{1}{\lambda^2}$.

Example 17.2. Returning to the lightbulb example, where $X \sim \mathrm{Exp}(2)$, we see that the average lifetime of a lightbulb is $\mathbb{E}X = \frac{1}{2}$ a year with variance $\mathrm{Var}(X) = \frac{1}{4}$.

If we wanted to calculate $\mathbb{P}(1 \leq X \leq 3)$, we could do this by integrating the PDF between 1 and 3. Alternatively, we could use the CDF:

$$\mathbb{P}(1 \leq X \leq 3) = F(3) - F(1) = (1 - \mathrm{e}^{-2\times 3}) - (1 - \mathrm{e}^{-2\times 1}) = \mathrm{e}^{-2} - \mathrm{e}^{-6} = 0.132.$$

Proof. of Theorem 17.1. For part 1,

$$\int_0^\infty \lambda \mathrm{e}^{-\lambda x}\,\mathrm{d}x = \big[-\mathrm{e}^{-\lambda x}\big]_0^\infty = -0 - (-1) = 1.$$

Similarly for part 2,

$$F(x) = \int_0^x \lambda \mathrm{e}^{-\lambda y} \, \mathrm{d}y = \big[- \mathrm{e}^{-\lambda y} \big]_0^x = -\mathrm{e}^{-\lambda x} - (-1) = 1 - \mathrm{e}^{-\lambda x}.$$

For part 3, we use integration by parts with $u = x$ and $v' = \lambda \mathrm{e}^{-\lambda x}$, so $u' = 1$ and $v = -\mathrm{e}^{-\lambda x}$. We get

$$\begin{aligned} \mathbb{E}X &= \int_0^\infty x\lambda \mathrm{e}^{-\lambda x} \, \mathrm{d}x \\ &= \big[- x\mathrm{e}^{-\lambda x} \big]_0^\infty + \int_0^\infty \mathrm{e}^{-\lambda x} \, \mathrm{d}x \\ &= -0 - (-0) + \left[-\frac{1}{\lambda} \mathrm{e}^{-\lambda x} \right]_0^\infty \\ &= -0 - \left(-\frac{1}{\lambda} \right) \\ &= \frac{1}{\lambda} \end{aligned}$$

For part 4, we use integration by parts with $u = x^2$ and $v' = \lambda \mathrm{e}^{-\lambda x}$, so $u' = 2x$ and $v = -\mathrm{e}^{-\lambda x}$ again. We get

$$\begin{aligned} \mathbb{E}X^2 &= \int_0^\infty x^2 \lambda \mathrm{e}^{-\lambda x} \, \mathrm{d}x \\ &= \big[- x^2 \mathrm{e}^{-\lambda x} \big]_0^\infty + \int_0^\infty 2x \mathrm{e}^{-\lambda x} \, \mathrm{d}x \\ &= -0 - (-0) + \frac{2}{\lambda} \int_0^\infty x \lambda x \mathrm{e}^{-\lambda x} \, \mathrm{d}x \\ &= \frac{2}{\lambda} \mathbb{E}X \\ &= \frac{2}{\lambda^2}, \end{aligned}$$

where we used a cunning trick on the integral on the right – spotting that we could turn it into the expectation, which is $1/\lambda$, by part 3 – to save us the effort of calculating it again. Hence

$$\operatorname{Var}(X) = \mathbb{E}X^2 - \left(\frac{1}{\lambda} \right)^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

□

17.2 Multiple continuous random variables

The theory we set up for two or more discrete random variables also works for two or more continuous random variables.

Now, the intensity of probability for (X, Y) being around (x, y) is given by the **joint probability density function** $f_{X,Y}$. In particular for $a \leq b$ and $c \leq d$, we have

$$\mathbb{P}(a \leq X \leq b \text{ and } c \leq Y \leq d) = \int_{x=a}^{b} \int_{y=c}^{d} f_{X,Y}(x, y) \, \mathrm{d}x \, \mathrm{d}y.$$

Discrete random variables	Continuous random variables
We can get the **marginal PMF** p_X of X by summing over y, so $p_X(x) = \sum_y p_{X,Y}(x, y)$.	We can get the **marginal PDF** f_X of X by integrating over y, so $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, \mathrm{d}y$.
Two discrete random variables X and Y are **independent** if their PMFs satisfy $p_{X,Y}(x, y) = p_X(x) \, p_Y(y) \qquad \text{for all } x, y$.	Two continuous random variables X and Y are **independent** if they have PDFs which satisfy $f_{X,Y}(x, y) = f_X(x) \, f_Y(y) \qquad \text{for all } x, y$.
The **conditional PMF** of Y given X is defined by $p_{Y\mid X}(y \mid x) = \frac{p_{X,Y}(x,y)}{p_X(x)}$.	The **conditional PDF** of Y given X is defined by $f_{Y\mid X}(y \mid x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$.
Bayes' theorem states that $p_{X\mid Y}(x \mid y) = \frac{p_X(x) \, p_{Y\mid X}(y\mid x)}{p_Y(y)}$.	**Bayes' theorem** states that $f_{X\mid Y}(x \mid y) = \frac{f_X(x) \, f_{Y\mid X}(y\mid x)}{f_Y(y)}$.
The expectation of a function of X and Y is given by the sum $\mathbb{E}g(X, Y) = \sum_{x,y} g(x, y) \, p_{X,Y}(x, y)$.	The expectation of a function of X and Y is given by the integral $\mathbb{E}g(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) \, f_{X,Y}(x, y) \, \mathrm{d}x \, \mathrm{d}y$.
The **covariance** of X and Y is given by $\mathrm{Cov}(X, Y) = \mathbb{E}(X - \mu_X)(Y - \mu_Y)$, and has a computational formula $\mathrm{Cov}(X, Y) = \mathbb{E}XY - \mu_X \mu_Y$.	The **covariance** of X and Y is given by $\mathrm{Cov}(X, Y) = \mathbb{E}(X - \mu_X)(Y - \mu_Y)$, and has a computational formula $\mathrm{Cov}(X, Y) = \mathbb{E}XY - \mu_X \mu_Y$.

Example 17.3. Consider the pair of continuous random variable (X, Y) with joint PDF

$$f_{X,Y}(x, y) = \tfrac{1}{2}(1 + x + y) \qquad \text{for } 0 \leq x, y \leq 1$$

and $f_{X,Y}(x, y) = 0$ otherwise.

We get the marginal distribution for X by integrating over y, so

$$f_X(x) = \int_0^1 \tfrac{1}{2}(1 + x + y) \, \mathrm{d}y = \tfrac{1}{2} \left[(1 + x)y + \tfrac{1}{2}y^2\right]_0^1 = \tfrac{3}{4} + \tfrac{1}{2}x.$$

We can find the the conditional PDF for Y given $X = \frac{1}{4}$. It is

$$f_{Y\mid X}(y \mid \tfrac{1}{4}) = \frac{f_{X,Y}(\frac{1}{4}, y)}{f_X(\frac{1}{4})} = \frac{\frac{1}{2}(1 + \frac{1}{4} + y)}{\frac{3}{4} + \frac{1}{2} \times \frac{1}{4}} = \tfrac{5}{7} + \tfrac{4}{7}y.$$

We can calculate the covariance. First, the expectations are

$$\mathbb{E}X = \int_{-\infty}^{\infty} x \, f_X(x) \, \mathrm{d}x = \int_0^1 x(\tfrac{3}{4} + \tfrac{1}{2}x) \, \mathrm{d}x = [\tfrac{3}{8}x^2 + \tfrac{1}{6}x^3] = \tfrac{13}{24}$$

and $\mathbb{E}Y = \frac{13}{24}$ also, by symmetry. Second, we have

$$\begin{aligned}
\mathbb{E}XY &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy\, f_{X,Y}(x,y)\, \mathrm{d}x\, \mathrm{d}y \\
&= \int_0^1 \int_0^1 xy\, \tfrac12 (1+x+y)\, \mathrm{d}x\, \mathrm{d}y \\
&= \int_0^1 \left[\tfrac14 x^2 y + \tfrac16 x^3 y + \tfrac14 x^2 y^2\right]_{x=0}^1 \mathrm{d}y \\
&= \int_0^1 \left(\tfrac14 y + \tfrac16 y + \tfrac14 y^2\right) \mathrm{d}y \\
&= \left[\tfrac18 y^2 + \tfrac1{12} y^2 + \tfrac1{12} y^3\right]_0^1 \\
&= \tfrac{7}{24}.
\end{aligned}$$

So therefore,

$$\mathrm{Cov}(X,Y) = \mathbb{E}XY - \mu_X \mu_Y = \tfrac{7}{24} - \tfrac{13}{24} \times \tfrac{13}{24} = -\tfrac{1}{576}.$$

Summary

- The exponential distribution has PDF $f(x) = \lambda \mathrm{e}^{-\lambda x}$, expectation $1/\lambda$, and variance $1/\lambda^2$.
- Most properties of multiple discrete random variables carry of to multiple continuous random variables. To get a marginal PDF from a joint PDF, we integrate (rather than sum) over the other variable.

Chapter 18

Limit theorems

There will be no lecture on Wednesday 30 November, as the UCU are on strike. Instead, you should learn the material that would have been given in the lecture from these notes. This material is fully examinable, as much as any other lecture.

I have provided some old pandemic-era videos embedded in these notes, which may help you – the section numbering has changed since these videos were made, but the material is mostly the same. Drop-in sessions are available on Thursday (2pm in PRD 9.320 and 4pm in Roger Stevens LT17) if you struggle with anything in this lecture.

18.1 Law of large numbers

In this lecture we will look at some "limit" results; that is, results about a large number of n of random variables, as n tends to infinity. We will be looking at the case when the random variables are independent and identically distributed (IID), which represents the case of multiple repeated experiments.

So 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 $\sigma^2 = \operatorname{Var}(X_1)$ for the common variance.

At the beginning of the course, we saw the mean of some values $x_1, x_2, \dots, x_n$ 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 $X_1, X_2, \dots, X_n$ 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 too. This mean random variable $\overline{X}_n$ will be out main object in this lecture. 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 used linearity of expectation to take the $1/n$ out of the brackets and to add up the individual expectations.

In the same way, the variance of $\overline{X}_n$ is

$$\begin{aligned}
\mathrm{Var}(\overline{X}_n) &= \mathrm{Var}\left(\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right) \\
&= \left(\frac{1}{n}\right)^2 \mathrm{Var}(X_1 + X_2 + \cdots + X_n) \\
&= \frac{1}{n^2}(\,\mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n)) \\
&= \frac{1}{n^2}(\sigma^2 + \sigma^2 + \cdots + \sigma^2) \\
&= \frac{1}{n^2}n\sigma^2 \\
&= \frac{\sigma^2}{n}.
\end{aligned}$$

Here, we crucially used the fact that the random variables are independent to write the variance of the sum as a sum of the variances.

In conclusion we have this:

Theorem 18.1. *Let $X_1, X_2, \dots, X_n$ be a sequence of IID random variables, and write $\mu = \mathbb{E}X_1$ for the common expectation and $\sigma^2 = \mathrm{Var}(X_1)$ for the common variance. Further, write $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ for the mean of these random variables. 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. So the range of probable values for $\overline{X}_n$ will be squeezing tighter and tighter around μ. Given that, it seems as if (and it can be rigorously proven that) we have the "law of large numbers".

Theorem 18.2 (Law of large numbers)**.** *Let $X_1, X_2, \dots$ 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*

$$\overline{X}_n \to \mu \quad \textit{in probability as } n \to \infty;$$

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 mean $\overline{X}_n$ is overwhelmingly likely to get closer and closer to the expectation $\mathbb{E}X = \mu$. In other words, the expectation $\mathbb{E}X = \mu$ represents the "long-run average" of independent experiments – this justifies referring to the expectation as a kind of average.

One special case is if we have repeated experiments that succeed with probability p; that is, $X_n \sim \mathrm{Bern}(p)$. 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 back to what we said about "frequentist probability" right at the beginning of Lecture 3: that one way to understand the probability of an event is as the long-run frequency of its occurrence.

18.2 Central limit theorem

We have seen that if the X_i are IID random variables with expectation μ and variance σ^2, then

$$\mathbb{E}\overline{X}_n = \mu \qquad \mathrm{Var}\left(\overline{X}_n\right) = \frac{\sigma^2}{n}$$

and the law of large numbers tells us that $\overline{X}_n \to \mu$ as $n \to \infty$. Alternatively, we could say that $\overline{X}_n - \mu \to 0$.

We might also want to know what the variation of $\overline{X}_n - \mu$ is around 0. Obviously, the law of large numbers tells us this variation shrinks away to 0, but we might be interested in the "shape" of that variation as it shrinks away. To study this variation, we need to "re-inflate" it, to stop it disappearing. It turns out that the correct way to do this is by multiplying by $\sqrt{n}$ and looking at $\sqrt{n}(\overline{X}_n - \mu)$.

It's fairly easy to check that the expectation and variance of the "re-inflated variation" is

$$\begin{aligned} \mathbb{E}\sqrt{n}(\overline{X}_n - \mu) &= \sqrt{n}(\mu - \mu) = 0 \\ \mathrm{Var}\left(\sqrt{n}(\overline{X}_n - \mu)\right) &= (\sqrt{n})^2\,\frac{\sigma^2}{n} = \sigma^2. \end{aligned}$$

So whatever distribution $\sqrt{n}(\overline{X}_n - \mu)$ has, that distribution must have expectation 0 and variance σ^2. But in fact, *no matter what distribution the X_i have* and regardless of whether they are discrete or continuous, this "variation around 0" $\sqrt{n}(\overline{X}_n - \mu)$ always gets closer and closer to the normal distribution!

Theorem 18.3 (Central limit theorem)**.** *Let $X_1, X_2, \ldots$ be a sequence of IID random variables. Write $\mu = \mathbb{E}X_1$ for the common expectation, $\sigma^2 = \mathrm{Var}(X_1)$ for the common variance, and $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ for the mean of the first n random variables. Then*

$$\sqrt{n}(\overline{X}_n - \mu) \to \mathrm{N}(0, \sigma^2) \quad \textit{in distribution as } n \to \infty;$$

by which we mean that, if $Y \sim \mathrm{N}(0, \sigma^2)$, then, for all $a < b$,

$$\mathbb{P}\left(a \leq \sqrt{n}(\overline{X}_n - \mu) \leq b\right) \to \mathbb{P}(a \leq Y \leq b) \quad \textit{as } n \to \infty.$$

(A full proof of the central limit theorem is too complicated to include here.)

Another alternative way to write this is to divide both sides by σ and write it as

$$\frac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}} \to \mathrm{N}(0, 1) \quad \text{in distribution as } n \to \infty.$$

The result we have stated, for IID random variables, is the most important case of the central limit theorem. But central limit theorems can be proved for other cases too – the rough principle is that if you have lots of random variables most of which are independent (or only weakly dependent) and none of which are individually too big, then the mean or sum will be approximately normally distributed.

18.3 Approximations with the normal distribution

There are many other distributions X that can be well approximated by a normal distribution where μ is set to $\mathbb{E}X$ and σ^2 is set to $\mathrm{Var}(X)$. Using intuition from the central limit theorem, this is roughly when the distribution can be expressed as the accumulation of many small effects.

- A binomial distribution $X \sim \mathrm{Bin}(n, p)$ is well approximated by a normal distribution $\mathrm{N}(np, np(1-p))$ when n is large and p is not too close to 0 or 1. (When p is small, we already know that the Poisson distribution is a good approximation.)
- A Poisson distribution $X \sim \mathrm{Po}(\lambda)$ is well approximated by a normal distribution $\mathrm{N}(\lambda, \lambda)$ when λ is large.
- A sum $Y = X_1 + \cdots + X_n$ of n IID geometric distributions $X_1, \ldots, X_n \sim \mathrm{Geom}(p)$ (sometimes known as a "negative binomial" distribution) is well approximated by a normal distribution $\mathrm{N}(n/p, np/(1-p)^2)$ when n is large and p is not to close to 1.

- A sum $Y = X_1 + \dots + X_n$ of n IID exponential distributions $X_1, \dots, X_n \sim \mathrm{Exp}(\lambda)$ (sometimes known as a "Gamma" distribution) is well approximated by a normal distribution $\mathrm{N}(n/\lambda, n/\lambda^2)$ when n is large and the expectation $1/\lambda$ is not too small.

We already know, of course, that a sum of independent normal distributions is *exactly* normal, with no approximations needed.

Example 18.1. *Suppose I toss 1000 coins. What's the probability I get between 495 and 505 Heads?*

The true distribution of Heads is $X \sim \mathrm{Bin}(1000, \frac{1}{2})$, and the question wants

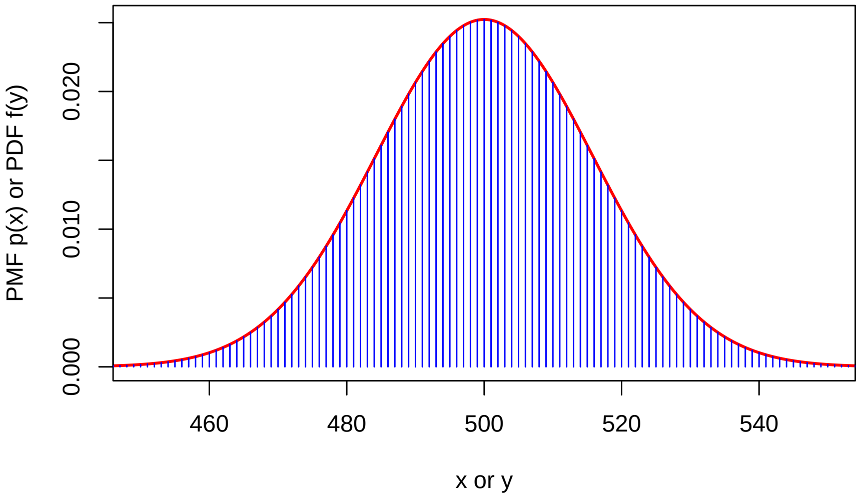
$$\mathbb{P}(495 \leq X \leq 505) = \sum_{x=495}^{505} p_X(x).$$

We can calculate the exact answer using R:

```
sum(dbinom(495:505, 1000, 1/2))
```

```
## [1] 0.2720284
```

However, we could instead use a normal approximation (which, again, would be useful in Victorian times or in an exam). Since $\mathbb{E}X = 1000 \times \frac{1}{2} = 500$ and $\mathrm{Var}(X) = 1000 \times \frac{1}{2} \times \frac{1}{2} = 250$, we have the normal approximation $Y \sim \mathrm{N}(500, 250)$. The figure below shows the PMF of the discrete distribution $X \sim \mathrm{Bin}(1000, \frac{1}{2})$ (blue bars) and the PDF of the continuous approximation $Y \sim \mathrm{N}(500, 250)$ (red line) between 450 and 550 – it is an extremely close match!



We could then calculate

$$\mathbb{P}(495 \leq X \leq 505) \approx \mathbb{P}(495 \leq Y \leq 505).$$

We could standardise and use the statistical tables, or just use R:

```
pnorm(505, 500, sqrt(250)) - pnorm(495, 500, sqrt(250))
```

```
## [1] 0.2481704
```

This is not too far off the correct answer 0.272 we calculated exactly, but it does miss by about 9%.

Note, though, that we approximated the discrete random variable X by a continuous random variable Y. So the next possibility for X above 505 was 506 and below 495 was 494, whereas Y could smoothly vary between the two. So we usually get a more accurate approximation if we use a **continuity correction** and round outwards halfway to the next discrete point. So we should get a better approximation from

$$\mathbb{P}(495 \leq X \leq 505) \approx \mathbb{P}(494.5 \leq Y \leq 505.5).$$

Calculating this in R (or with statistical tables) we get

```
pnorm(505.5, 500, sqrt(250)) - pnorm(494.5, 500, sqrt(250))
```

```
## [1] 0.2720476
```

Using the continuity correction, we now have an incredibly accurate approximation – it only misses by 0.006%.

Whenever we approximate a discrete random variable by a continuous random variable (such as a normal distribution), using a continuity correction – that is, rounding halfway to the next discrete point – typically makes the approximation more accurate. If X is a discrete random variable that takes integer values and Y is a continuous approximation, then the appropriate continuity corrections are

$$\begin{aligned} \mathbb{P}(X \leq n) &\approx \mathbb{P}(Y \leq n + \tfrac{1}{2}) &\qquad \mathbb{P}(X \geq n) &\approx \mathbb{P}(Y \geq n - \tfrac{1}{2}) \\ \mathbb{P}(X < n) &\approx \mathbb{P}(Y < n - \tfrac{1}{2}) &\qquad \mathbb{P}(X > n) &\approx \mathbb{P}(Y > n + \tfrac{1}{2}) \end{aligned}$$

Summary

- If $\overline{X}_n$ is the mean of n IID random variables with expectation μ and variance σ^2, then $\mathbb{E}\overline{X}_n = \mu$ and $\operatorname{Var}(\overline{X}_n) = \sigma^2/n$.
- The law of large numbers says that $\overline{X}_n \to \mu$.
- The central limit theorem says that $\sqrt{n}(\overline{X}_n - \mu) \to \mathrm{N}(0, \sigma^2)$.

Problem Sheet 5

This is Problem Sheet 5. This problem sheet covers Lectures 15 to 18. You should work through all the questions on this problem sheet in preparation for your tutorial in Week 10. The problem sheet contains two assessed questions, which are due in by 2pm on Monday 12 December.

A: Short questions

A1. Consider the continuous random variable X with PDF

$$f(x) = \begin{cases} \frac12 x & \text{for } 0 \leq x \leq 1 \\ \frac12 & \text{for } 1 < x \leq 2 \\ \frac32 - \frac12 x & \text{for } 2 < x \leq 3 \end{cases}$$

and $f(x) = 0$ otherwise.

(a) Calculate the CDF for X.

(b) What is $\mathbb{P}(\frac32 \leq X \leq \frac52)$?

(c) Calculate the expectation $\mathbb{E}X$.

A2. Let X be a continuous random variable with PDF

$$f(x) = \frac{k}{x^3} \qquad \text{for } x \geq 1$$

and $f(x) = 0$ otherwise.

(a) What value of k makes this into a true PDF?

(b) What is $\mathbb{P}(X \geq 3)$?

(c) What is the expected value $\mathbb{E}X$?

A3. Let $X \sim \mathrm{Exp}(\frac12)$.

(a) What is $\mathbb{E}X$?

(b) What is $\mathbb{P}(1 \leq X \leq 3)$?

A4. Let $Z \sim \mathrm{N}(0, 1)$. Calculate the following **(a)** using statistical tables; **(b)** using R. (For part (a), you should show enough working to convince a reader that you really did use the tables.)

(i) $\mathbb{P}(Z \leq -1.2)$

(ii) $\mathbb{P}(-1.2 \leq Z \leq 0.8)$

(iii) $\mathbb{P}(Z \leq 0.27)$ (using interpolation for part (a))

A5. Let $X \sim \text{Po}(25)$. Calculate the following **(a)** exactly, using R; **(b)** approximately, using a normal approximation with a continuity correction and statistical tables. (For part (b), you should show enough working to convince a reader that you really did use the tables.)

(i) $\mathbb{P}(X \leq 27)$

(ii) $\mathbb{P}(X \geq 28 \mid X \geq 27)$

B: Long questions

B1. **(a)** Let $X \sim \text{Exp}(\lambda)$. Show that

$$\mathbb{P}(X > x + y \mid X > y) = \mathbb{P}(X > x).$$

(b) The result proved in part (a) is called the "memoryless property". Why do you think it's called that?

(c) When you get to certain bus stop, the average amount of time you have to wait for a bus to arrive is 20 minutes. Specifically, the time until the next bus arrives is modelled as an exponential distribution with expectation $1/\lambda = 20$ minutes. Suppose you have already been waiting at the bus stop for 15 minutes. What is the expected further amount of time you still have to wait for a bus to arrive?

B2. The main dangerous radioactive material left over after the Chernobyl disaster is Caesium-137. The amount of time it takes a Caesium-137 particle to decay is known to follow an exponential distribution with rate $\lambda = 0.023$ years^{-1}.

(a) What is the average amount of time it takes a Caesium-137 particle to decay?

(b) The "half-life" of a radioactive substance is the amount of time it takes for half of the substance to decay. Using the information in the question, calculate the half-life of Caesium-137.

(c) It is estimated that roughly 24 kg of Caesium-137 was released during the Chernobyl disaster, which happened roughly 35.6 years ago. Estimate the mass of Caesium-137 that has still not decayed?

B3. Consider the pair of random variables (X, Y) with joint PDF

$$f_{X,Y}(x, y) = 2 \qquad \text{for } 0 \leq x \leq y \leq 1$$

and $f_{X,Y}(x, y) = 0$ otherwise. (In particular, note that the joint PDF is only nonzero when $x \leq y$.)

(a) Draw a picture of the range of (X, Y) in the xy-plane.

(b) Describe the conditional distribution of X given $Y = y$, for $0 \leq y \leq 1$.

(c) What is the marginal PDF f_X of X?

(d) Are X and Y independent?

B4. *(Optional)* Engineers and scientists often use the rule of thumb "Only 5% of data is more than two sample standard deviations away from the sample mean." Carefully justify this rule, using concepts from the module.

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 may find it helpful to recognise some expectations as definitional formulas for variances, where appropriate.)

(c) At the beginning of this module, 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 $1/n$.

B6. *(New)* Roughly how many times should I toss a coin for there to be a 95% chance that between 49% and 51% of my coin tosses land Heads?

(I meant to delete Question B4 from this problem sheet, to make it shorter, but I accidentally deleted Question B6 instead. I've now added B6 and marked B4 as "optional".)

C: Assessed questions

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

The deadline for submitting your solutions is **2pm on Monday 12 December** at the beginning of Week 11. Submission will be via Gradescope; submission will open on Monday 5 November. Your work will be marked by your tutor and returned later, 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 X be a continuous random variable with PDF

$$f(x) = \tfrac{2}{9}(2-x) \qquad \text{for } -1 \leq x \leq c$$

and $f(x) = 0$ otherwise.

(a) Explaining your work, find the value of the constant c.

(b) What is $\mathbb{P}(X > 1)$?

(c) Calculate the expectation of$X$.

(d) Calculate the variance of X.

C2. For each of the following, **(a)** calculate the *exact* value using R; **(b)** get an approximate value using an appropriate approximation and *without* using R. (Statistical tables are available.)

(i) $\mathbb{P}(X \leq 3)$, where $X \sim \text{Bin}(1000, 0.005)$.

(ii) $\mathbb{P}(296 \leq Y \leq 307)$, where $Y \sim \text{Bin}(1200, 0.25)$.

(iii) $\mathbb{P}(Z \geq 398)$, where $Z \sim \text{Bin}(400, 0.995)$.

Solutions to short questions

A1. (b) $\frac{7}{16}$ (c) 1.5
A2. (a) 2 (b) $\frac{1}{9}$ (c) 2
A3. (a) 2 (b) 0.383
A4. (i) 0.1150 (ii) 0.6731 (or 0.6730 with tables) (iii) 0.6064
A5. (a) (i) 0.7002 (ii) 0.809 (b) (i) 0.6914 (ii) 0.807

Part III: Bayesian statistics

Chapter 19

The Bayesian idea

In this final section of the module, we return to statistics, where we will look at an approach to data analysis known as "Bayesian statistics".

Statistics concerns how to draw conclusions from data; and **Bayesian statistics** is one particular framework for doing this. The idea of Bayesian statistics is that we start with "prior" ("before") beliefs about the underlying model, then use the data (together with Bayes' theorem) to update our to our "posterior" ("after") beliefs about the model *given* the data we have observed.

19.1 Example: fake coin?

We will start by illustrating the main idea with an example.

Example 19.1. *A joke shop sells three types of coins: normal fair coins; Heads-biased coins, which land Heads with probability 0.8; and Tails-biased coins, which land Heads with probability 0.2. I pick up a coin and examine it; since it looks mostly like a normal coin, I believe there's 60% chance it's s fair coin, and a 20% chance it's biased either way. I decide to toss the coin four times, to gather some more evidence. The result is that all three are Heads. How should I update my beliefs?*

So, we start with the "prior" (before) belief

$$\mathbb{P}(\text{fair}) = 0.6 \qquad \mathbb{P}(\text{H-bias}) = 0.2 \qquad \mathbb{P}(\text{T-bias}) = 0.2$$

We know how to update our beliefs after seeing the data: we use Bayes' theorem. We have

$$\begin{aligned}
\mathbb{P}(\text{fair} \mid \text{HHH}) &= \frac{\mathbb{P}(\text{fair})\,\mathbb{P}(\text{HHH} \mid \text{fair})}{\mathbb{P}(\text{HHH})} = \frac{0.6 \times 0.5^3}{\mathbb{P}(\text{HHH})} = \frac{0.075}{\mathbb{P}(\text{HHH})} \\
\mathbb{P}(\text{H-bias} \mid \text{HHH}) &= \frac{\mathbb{P}(\text{H-bias})\,\mathbb{P}(\text{HHH} \mid \text{H-bias})}{\mathbb{P}(\text{HHH})} = \frac{0.2 \times 0.8^3}{\mathbb{P}(\text{HHH})} = \frac{0.1024}{\mathbb{P}(\text{HHH})} \\
\mathbb{P}(\text{T-bias} \mid \text{HHH}) &= \frac{\mathbb{P}(\text{H-bias})\,\mathbb{P}(\text{HHH} \mid \text{T-bias})}{\mathbb{P}(\text{HHH})} = \frac{0.2 \times 0.2^3}{\mathbb{P}(\text{HHH})} = \frac{0.0016}{\mathbb{P}(\text{HHH})}.
\end{aligned}$$

We also need to find common denominator $\mathbb{P}(\text{HHH})$. We could do that using the law of total probability. But a convenient short-cut is to notice that the above three probabilities have to add up to 1, and so that common denominator must be \$0.075 + 0.1024 + 0.0016 = 0.179, so we have

$$\begin{aligned}\mathbb{P}(\text{fair} \mid \text{data}) &= \frac{0.075}{0.179} = 0.419 \\ \mathbb{P}(\text{H-bias} \mid \text{data}) &= \frac{0.1024}{0.179} = 0.572 \\ \mathbb{P}(\text{T-bias} \mid \text{data}) &= \frac{0.0016}{0.179} = 0.009.\end{aligned}$$

So, after tossing the coin four times, our belief has been updated from the "prior" (before) belief

$$\mathbb{P}(\text{fair}) = 0.6 \qquad \mathbb{P}(\text{H-bias}) = 0.2 \qquad \mathbb{P}(\text{T-bias}) = 0.2$$

to the "posterior" (after) belief

$$\mathbb{P}(\text{fair} \mid \text{data}) = 0.419 \qquad \mathbb{P}(\text{H-bias} \mid \text{data}) = 0.572 \qquad \mathbb{P}(\text{T-bias} \mid \text{data}) = 0.009.$$

Compared to our prior beliefs, our belief that the coin is fair has decreased a little bit; our belief the coin is biased towards Heads has shot up, and is now our most likely belief; while our belief the coin is biased towards Tails has plummeted to just 1%.

19.2 Bayesian framework

Let's think more systematically about what we did in the previous example.

- **Model:** The four coin tosses were modelled as four IID Bernoulli trials $X_1, X_2, X_3, X_4 \sim \text{Bern}(\theta)$ (if we let $X_i = 1$ denote that the ith coin was Heads). Here, the probability of Heads is some unknown parameter θ. (Recall we talked about parametric models for data in Subsection 11.3.) This model gives a distribution that depends on the parameter. Hre we had a conditional PMF for one trial

  $$p(x \mid \theta) = \theta^x (1-\theta)^{1-x}$$

  (this is a convenient way of writing the PMF for a Bernoulli trial). So the joint PMF for the IID trials is

  $$p(\mathbf{x} \mid \theta) = \prod_{i=1}^{4} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_i x_i}(1-\theta)^{4-\sum_i x_i}.$$

  (Here and throughout, $\prod$, the Greek capital Pi, means a product – it's the multiplication equivalent of the summation Sigma, Σ.)
- **Prior:** We started with a prior belief $\pi(\theta)$ on the value of the unknown parameter. In our case, we had the PMF

  $$\pi(0.2) = 0.2 \qquad \pi(0.5) = 0.6 \qquad \pi(0.8) = 0.2.$$

- **Data:** We collected the data $\mathbf{x}$, which here had $x_1 = 1$, $x_2 = 1$, $x_3 = 1$ (with 1 denoting Heads and 0 denoting Tails).
- **Posterior:** We calculated the posterior distribution $\pi(\theta \mid \mathbf{x})$ for the parameter *given* the data. We did this using Bayes' theorem:

$$\pi(\theta \mid \mathbf{x}) = \frac{\pi(\theta)\, p(\mathbf{x} \mid \theta)}{p(\mathbf{x})} \propto \pi(\theta)\, p(\mathbf{x} \mid \theta).$$

  We recovered the constant of proportionality – that is, the denominator of Bayes' theorem – because we knew $\pi(\theta \mid \mathbf{x})$ was a conditional PMF so must add up to 1. We ended up with

$$\pi(0.2 \mid \mathbf{x}) = 0.009 \qquad \pi(0.5 \mid \mathbf{x}) = 0.419 \qquad \pi(0.8 \mid \mathbf{x}) = 0.572.$$

This is the framework of how Bayesian statistics works: model, prior, data, posterior. To lay it out more generally, the procedure goes like this:

- **Model:** We start with a model for the data $\mathbf{x}$ that depends on one or more parameters θ, as expressed by a conditional PMF (for discrete data) or PDF (for continuous data) $p(\mathbf{x} \mid \theta)$. This normally represents n IID experiments, so

$$p(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} p(x_i \mid \theta).$$

  This conditional distribution is often called the **likelihood**.
- **Prior:** We have a prior distribution $\pi(\theta)$ for the parameter θ, which can be either a PMF or PDF. The prior distribution represents our beliefs about the parameter before we collect the data; this can be based on previous evidence, expert opinion, personal intuition, etc.
- **Data:** We collect the data $\mathbf{x}$.
- **Posterior:** We then form the posterior distribution $\pi(\theta \mid \mathbf{x})$ for the parameter given the data, using Bayes' theorem:

$$\begin{gathered}\pi(\theta \mid \mathbf{x}) \propto \pi(\theta)\, p(\mathbf{x} \mid \theta) \\ \text{posterior} \propto \text{prior} \times \text{likelihood}.\end{gathered}$$

  This can either be a conditional PMF or PDF, but will be the same type as the prior $\pi(\theta)$.

19.3 Normal–normal model

In our first example, the "joke coins" was a bit artificial, giving us a prior with only three points in its range. Its often more appropriate to have a prior distribution for a parameter that is continuous over a wide range of possibilities (although concentrated towards the more parameters values we believe are more probable.).

Consider a normal likelihood, where $X_1, X_2, \dots, X_n$ are IID $\mathrm{N}(\theta, \sigma^2)$, and where the expectation θ is the unknown parameter but the variance σ^2 is known. This could model trying to measure some quantity θ with an instrument which is known to have a $\mathrm{N}(0, \sigma^2)$ error. So the model has joint PDF

$$p(\mathbf{x} \mid \theta) \propto \prod_{i=1}^{n} \exp\left(-\frac{(x_i - \theta)^2}{2\sigma^2}\right) \qquad = \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\theta)^2}{\sigma^2}\right)$$
$$= \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta)^2\right).$$

(Again, we only worry about distributions up to proportionality, because we work out the multiplicative constant at the end.)

In fact, when doing Bayesian statistics, it's often convenient to write $\tau = 1/\sigma^2$ for the inverse of the known variance; this τ is called the **precision** and is also known. So with this notation, the model is that $X_1, X_2, \dots, X_n$ are IID $\mathrm{N}(\theta, 1/\tau^2)$, with joint PDF

$$p(\mathbf{x} \mid \theta) \propto \exp\left(-\frac{\tau}{2}\sum_{i=1}^{n}(x_i - \theta)^2\right).$$

What about a prior for the unknown expectation θ. Often an appropriate choice is a normal $\mathrm{N}(\mu_0, 1/\tau_0)$ prior for the unknown expectation parameter θ. This represents that we expect the quantity we are trying to mention to be around μ_0, with an amount of uncertainty captured by the precision τ_0 on the prior. So the prior PDF is

$$\pi(\theta) \propto \exp\left(-\frac{\tau_0}{2}(\theta - \mu_0)^2\right)$$

Because both the prior distribution and the model for the data are normal, this is known as the **normal–normal model**.

Suppose we collect data $\mathbf{x} = (x_1, x_2, \dots, x_n)$, and recall that we write $\bar{x} = (\sum_i x_i)/n$ for the sample mean.

To get the posterior distribution requires a bit of an algebra slog (see below), but the outcome is that the posterior distribution is

$$\pi(\theta \mid \mathbf{x}) \propto \exp\left(-\frac{\tau_0 + n\tau}{2}\left(\theta - \frac{\tau_0\mu_0 + n\tau\bar{x}}{\tau_0 + n\tau}\right)^2\right),$$

which is (proportional to) the PDF for yet another normal distribution

$$\theta \mid \mathbf{x} \sim \mathrm{N}\left(\frac{\tau_0}{\tau_0 + n\tau}\mu_0 + \frac{n\tau}{\tau_0 + n\tau}\bar{x},\ \frac{1}{\tau_0 + n\tau}\right).$$

In other words, the posterior expectation

$$\mathbb{E}(\theta \mid \mathbf{x}) = \frac{\tau_0}{\tau_0 + n\tau}\mu_0 + \frac{n\tau}{\tau_0 + n\tau}\bar{x}$$

is a weighted average of the prior expectation μ_0 and the mean of the data $\bar{x}$, and the more datapoints n you get, the heavier the weighting on the data compared to the prior. Further, the precision has increased from the prior precision τ_0 to the posterior precision $\tau_0 + n\tau$; so the more data we get, the larger the precision gets, so the smaller the variance gets, and the more sure we get about the true value of θ.

The algebra slog (non-examinable). Recall that we can ignore multiplicative terms that don't contain θ, thanks to our proportionality trick. But note also that a multiplicative term becomes an additive term inside an exponential. So, within an exponential, we can always ignore any "plus constants" that don't involve θ.

So the prior can be written as

$$\begin{aligned}\pi(\theta) &\propto \exp\left(-\frac{\tau_0}{2}(\theta-\mu_0)^2\right)\\ &= \exp\left(-\frac{\tau_0}{2}\theta^2 + \tau_0\mu_0\theta - \frac{\tau_0}{2}\mu_0^2\right)\\ &\propto \exp\left(-\frac{\tau_0}{2}\theta^2 + \tau_0\mu_0\theta\right),\end{aligned}$$

where we ignored the final constant term.

Similarly, the likelihood can be written as

$$\begin{aligned}p(\mathbf{x}\mid\theta) &= \exp\left(-\frac{\tau}{2}\sum_{i=1}^n (x_i-\theta)^2\right)\\ &= \exp\left(-\frac{\tau}{2}\sum_{i=1}^n x_i^2 + \tau\theta\sum_{i=1}^n x_i - \frac{\tau}{2}n\theta^2\right)\\ &\propto \exp\left(n\tau\theta\bar{x} - \frac{\tau}{2}n\theta^2\right),\end{aligned}$$

where we ignored the first constant term in the second line, and recognised $\sum_i x_i$ as $n\bar{x}$, as we have done before.

Then Bayes' theorem gives us

$$\begin{aligned}\pi(\mathbf{x}\mid\theta) &\propto \pi(\theta)\,p(\mathbf{x}\mid\theta)\\ &\propto \exp\left(-\frac{\tau_0}{2}\theta^2 + \tau_0\mu_0\theta\right)\times\exp\left(n\tau\theta\bar{x} - \frac{\tau}{2}n\theta^2\right)\\ &= \exp\left(-\frac{\tau_0}{2}\theta^2 + \tau_0\mu_0\theta + n\tau\theta\bar{x} - \frac{\tau}{2}n\theta^2\right)\\ &= \exp\left(-\frac{\tau_0+n\tau}{2}\theta^2 + (\tau_0\mu_0 + n\tau\bar{x})\theta\right)\\ &= \exp\left(-\left(\frac{\tau_0+n\tau}{2}\right)\left(\theta^2 - 2\frac{\tau_0\mu_0+n\tau\bar{x}}{\tau_0+n\tau}\theta\right)\right)\\ &\propto \exp\left(-\tfrac{\tau_0+n\tau}{2}\left(\theta - \frac{\tau_0\mu_0+n\tau\bar{x}}{\tau_0+n\tau}\right)^2\right).\end{aligned}$$

In the final line, the squared term differs from the line above only by some additive constant. But this is exactly (proportional to) the PDF of a normal distribution with expectation

$$\frac{\tau_0\mu_0+n\tau\bar{x}}{\tau_0+n\tau}$$

and precision $\tau_0 + n\tau$.

Chapter 20

More Bayesian models

20.1 Beta distribution

In our fake-coin example in the last lecture, we had a prior PMF for the parameter $\theta = p$ that could only take one of three possible values. But when doing Bayesian statistics with a parameter that represents a probability, it makes more sense to have a prior PDF that covers the whole interval $[0, 1]$. After all, any parameter value that is given a probability of 0 in the prior always has a probability 0 in the posterior as well, no matter how strong the evidence in its favour; it's considered good practice to only put 0 prior probability on parameter values that are *literally impossible*, such as probabilities below 0 or above 1. (This is sometimes called "Cromwell's rule".)

One useful family of distributions to use as a prior distribution for a probability parameter is the Beta distribution, whose range is the whole interval $[0, 1]$.

Definition 20.1. A continuous random variable X is said to have the **Beta distribution** with parameters α and β if it has the PDF

$$f(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1} \qquad \text{for } 0 \leq x \leq 1$$

and 0 otherwise. Here, the constant

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,\mathrm{d}x,$$

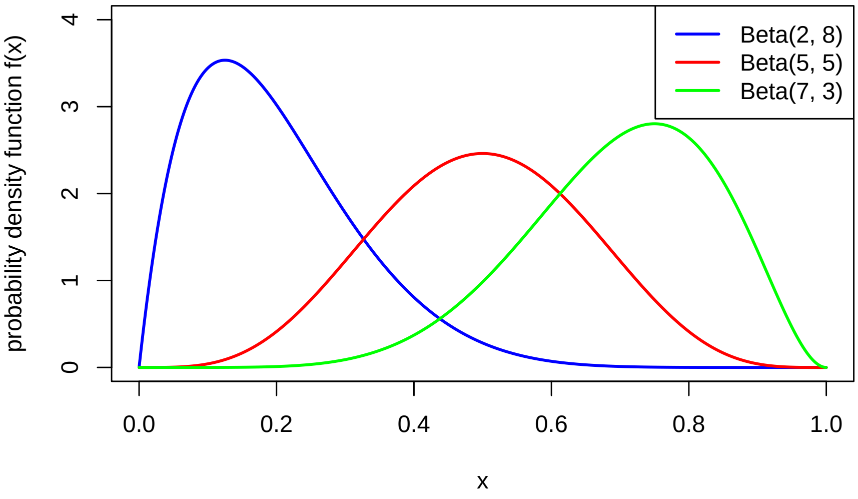
known as the "Beta function", ensures that the PDF integrates to 1. We write $X \sim \mathrm{Beta}(\alpha, \beta)$.

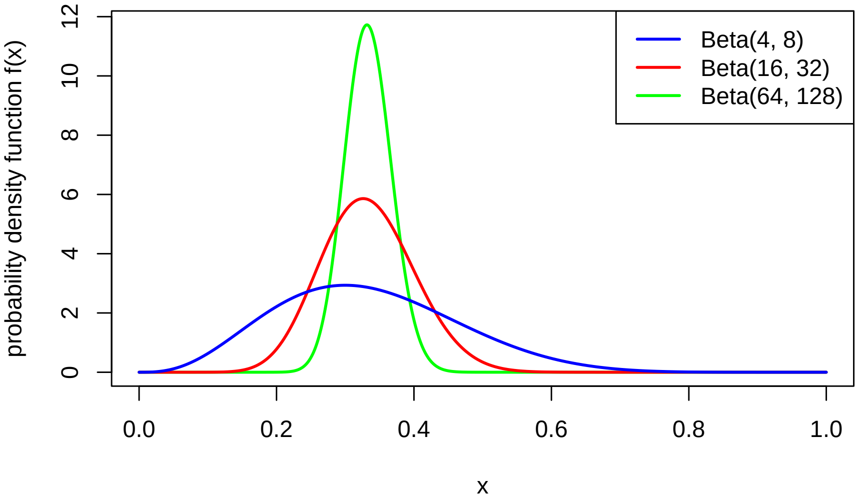
Theorem 20.1. *Let $X \sim Beta(\alpha, \beta)$. Then*

1. $\mathbb{E}X = \dfrac{\alpha}{\alpha+\beta}$
2. $\mathrm{Var}(X) = \dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \dfrac{\mu(1-\mu)}{\alpha+\beta+1}$, *where* $\mu = \mathbb{E}X$.

(Proving this requires some awkward messing around with Gamma functions, which we won't bother with here.)

So the idea is that the expectation of X is decided on by the *relative* values of α and β, while the variance is decided by the *total* value of α and β. The following two pictures illustrate this:





Note also that Beta$(1, 1)$ is the continuous uniform distribution from Example 15.1.

Example 20.1. *A statistician is studying the probability θ that ordinary coins land Heads. She would like to use a prior distribution for θ with prior expectation* 0.5 *and prior standard deviation* 0.01*. What Beta distribution would be appropriate to use?*

To get $\mathbb{E}\theta = 0.5$, we need $\alpha = \beta$. Then the variance, which needs to be $0.01^2 = 0.0001$, is

$$\operatorname{Var}(\theta) = \frac{\mu(1-\mu)}{\alpha+\beta+1} = \frac{0.25}{\alpha+\beta+1}.$$

This requires $\alpha = \beta = 1250$. (Well, actually 1249.5.)

20.2 Beta–Bernoulli model

Consider a Bernoulli likelihood, where $X_1, X_2, \dots, X_n$ are IID $\mathrm{Bern}(\theta)$, so have joint PMF

$$p(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i} = \theta^{y}(1-\theta)^{n-y},$$

where we have written $y = \sum_i x_i$ for the total number of successes. Consider further using a $\mathrm{Beta}(\alpha, \beta)$ prior for θ, so that

$$\pi(\theta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

(Because we're going to use the "posterior has to add up to 1" trick at the end, we're free to drop constants whenever we want.) This is known as the **Beta–Bernoulli model**.

Suppose we collect data $\mathbf{x} = (x_1, x_2, \dots, x_i)$, with $y = \sum_i x_i$ successes. What now is the posterior distribution for θ given this data?

Using Bayes' theorem, we have

$$\begin{aligned}\pi(\mathbf{x} \mid \theta) &\propto \pi(\theta)p(\mathbf{x} \mid \theta) \\ &= \theta^{\alpha-1}(1-\theta)^{\beta-1} \times \theta^{y}(1-\theta)^{n-y} \\ &= \theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1}.\end{aligned}$$

We can recognise immediately that this is proportional to the PDF for a $\mathrm{Beta}(\alpha + y, \beta + n - y)$ distribution, so in particular, the constant of proportionality must be $1/B(\alpha + y, \beta + n - y)$.

So we see that, like the prior, the posterior is also a Beta distribution, where the first parameter has gone from α to $\alpha + y$ and the second parameter has gone from β to $\beta + (n - y)$. In other words, α has increased by the number of successes, and β has increased by the number of failures. The expectation has gone from the prior expectation

$$\frac{\alpha}{\alpha+\beta}$$

to the posterior expectation

$$\frac{\alpha + y}{\alpha + \beta + n}.$$

This can be thought of as a sort of average between the prior expectation $\alpha/(\alpha+\beta)$ and the mean of the data y/n.

20.3 Modern Bayesian statistics

In this section, we've given just a brief taster of Bayesian statistics. Bayesian statistics is a deep and complicated subject, and you may have the opportunity to find out a lot more about it later in your university career.

We have seen that in Bayesian statistics, one brings in a subjective "prior" based on previous beliefs and evidence, then updates this prior based on the data. This contrasts with the more traditional **frequentist statistics**. In frequentist one uses only the data – no prior beliefs! – and judges to what extent the data is consistent or inconsistent with a hypothesis, without weighing in on how likely such a hypothesis is. (Frequentist statistics is the main subject studied in MATH1712 Probability and Statistics II.)

In the two main examples of Bayesian statistics we have looked at – the Bernoulli likelihood and the normal likelihood – we ended up with a posterior in the same parametric family as prior, just with different parameters. Such a prior is called a "conjugate prior". Of course, these are very convenient and easy to work with. However, with more complicated likelihoods and more complicated priors – especially those not with a single parameter but with many parameters – calculating the posterior distribution can be very difficult. In particular, working out the constant of proportionality (even just approximately) and/or sampling from the posterior distribution are very hard problems.

For this reason, Bayesian statistics was for a long time a minor area of statistics. However, increases in computer power in the 1980s made some of these problems more tractable, and Bayesian statistics has increased in importance and popularity since then.

For a while, there was an occasionally fierce debate between "Bayesians" and "frequentists". Frequentists thought that bringing subjective personal beliefs into things was unmathematical, while Bayesians thought that ignoring how plausible a hypothesis is before testing it is unscientific. The debate has now largely dissipated, and it is largely accepted that modern statisticians need to know about both frequentist and Bayesian methods.

There are still plenty of open problems in Bayesian statistics, and lots of these involve the computational side: finding algorithms that can efficiently calculate the normalising constants in posterior distributions or sample from those posterior distributions, especially when the parameter(s) have very high dimension.

Summary

- In Bayesian statistics, we start with a prior distribution for a parameter θ, and update to a posterior distribution given the data $\mathbf{x}$, through $\pi(\theta \mid \mathbf{x}) \propto \pi(\theta)p(\mathbf{x} \mid \theta)$, or posterior $\propto$ prior $\times$ likelihood.
- The Beta distribution is a useful family of distributions to use as priors for probability parameters.
- A Beta prior for a Bernoulli likelihood leads to a Beta posterior with different parameters.
- A normal prior for the expectation of a normal likelihood wioth known variance leads to a normal posterior with different parameters.

Problem Sheet 6

This is not a proper problem sheet, but is a few questions on Bayesian statistics to test your knowledge of Section 10. There is no assessed work on this sheet.

1. I want to use a prior distribution for a parameter θ whose range is the interval $[0, 1]$, whose expectation is 0.4 and whose standard deviation is 0.2. Suggest an appropriate distribution.

2. My data is modelled as having a $\mathrm{Bern}(\theta)$ likelihood, and I plan to record 10 IID observations. I choose to use a $\mathrm{Beta}(1, 4)$ prior.

(a) What is the prior expectation and variance?

(b) Suppose my data records 2 successes and 8 failures. What is the posterior expectation and variance?

(c) Suppose my data records 5 successes and 5 failures. What is the posterior expectation and variance?

(d) Briefly comment on these results.

3. **(a)** My data is modelled as a single data point with a $\mathrm{Geom}(\theta)$ likelihood, so

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

I use a $\mathrm{Beta}(\alpha, \beta)$ prior for θ. Show that the posterior distribution is $\mathrm{Beta}(\alpha + 1, \beta + x - 1)$.

(b) I instead choose to collect n IID data points, using the same geometric likelihood and Beta prior. Show that the posterior distribution is a Beta distribution, and state the parameters.

(c) Compare your results to that of the Beta–Bernoulli model, and briefly comment.

Other stuff

R Worksheets

R worksheets

Each week there will be an R worksheet to work through in your own time. I recommend spending about one hour on each worksheet, plus one extra hour for worksheets with assessed questions, for checking and submitting your solutions.

Week	Worksheet	Deadline for assessed work
1	**R basics** (Solutions)	—
2	**Vectors**	—
3	**Data in R**	Monday 24 October (Week 4)
4	**Plots I:** Making plots	—
5	**Plots II:** Making plots better	Monday 7 November (Week 6)
6	**RMarkdown** (optional) [Rmd]	—
7	**Discrete distributions** [Rmd]	Monday 21 November (Week 8)
8	**Discrete random variables** [Rmd]	—
9	**Normal distribution** [Rmd]	Monday 5 December (Week 10)
10	**Law of large numbers** [Rmd]	—
11	**Recap**	Thursday 15 December (Week 11)

About R and RStudio

- **R** is a *programming language* that is particularly useful for 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 modules 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 few ways you can access R and RStudio.

First, you can **install R and RStudio on your own computer**. Students who have their own computer (with administration and installation rights) usually find this the most convenient way use R.

When you install R and RStudio, it's important that you install R (the programming language) first, and only install RStudio (the program to use R) after R has already been installed. This ensures that RStudio can "find" R on your computer.

1. *First*, install R. Go to the Comprehensive R Archive Network (CRAN) 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 typically older Intel-based Macbooks, you must use the "Intel 64-bit build"; for post-November 2020 M1 or M2-based "Apple silicon" Macbooks, the "Apple silicon arm64 build" may be slightly 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 R troubleshooting 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.

Second, you can **use R and RStudio on University computers**. All University computers have access to R and RStudio, via the AppsAnywhere service. *First*, launch R via AppsAnywhere (at the time of writing, it's confusingly named "Cran R 4.2.0 x64"). You can then close the program that opens. *Then* launch RStudio ("Rstudio 2022"), also via AppsAnywhere. (You can decline any updates that are suggested.)

The R drop-in sessions take place in computer rooms, so if you have problems accessing R and RStudio on University computers, you can get help at the drop-in sessions too.

Third, you can **use the RStudio Cloud**. The RStudio Cloud is a cloud-hosted "Google Docs for R" that you can use through your web browser, without having to install anything. You can get 25 hours per month for free, which should be plenty for this module, or pay for more.

If you have access to a computer on which you can't install software, such as some Chromebooks or tablet computers, or if you're borrowing a friend's laptop, the RStudio Cloud can be a convenient solution.

Update: If you have an Intel-based Chromebook, then we have had success installing R and RStudio using these (rather complicated) instructions, although you may still find it more convenient just to use the RStudio Cloud.

R 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 difficult. Also, sometimes students have problems installing R and RStudio on their own computers. To help with this, we have organised optional R troubleshooting drop-in sessions in Weeks 2 and 3. Check your timetable for details – they are probably listed as "computer practicals".

Tips on writing mathematics

In the mid-semester survey, a few people suggested I could offer some clearer advice on writing mathematics well. This is a brief attempt at doing that. I may try to expand this a bit later.

Advice

If a maths question asks you to “State”, “Write down” or “Calculate” something, you don’t need to do any more than give the answer. But if a question asks you to “Prove”, “Show that” or “Explain”, then the marker is looking for a clearly explained answer, and will base your mark on how well you explain your solution, not just on it’s mathematical accuracy.

Some things a marker might look for include:

- If you had to make your own notation, have you explained clearly what it means?
- If it’s a “words question”, have you clearly translated the information into mathematical notation?
- In your solution, do you clearly state how you know certain statements are true? (“From the question, we know that…” “From the definition of …” “Using the law of total probability…” “From Axiom 2, we see that…”)
- When making non-obvious algebraic manipulations, do you explain what you’re doing? (There’s usually no need – but no harm either! – in stating obvious things like “Dividing both sides by 2…”, but for less obvious things like “Recognising this as difference of two squares…” or “By the linearity of expectation…”, you should say so.)
- Does your solution show clearly the “direction” of the proof. (“We start with the definition of…” “We need to show that…” “Assume, seeking a contradiction, that…” “If we could show that … then this would be sufficient to give the result.” “We will bound each term in this expression separately.”)
- Do you write in full sentences? Even equations should (usually) include punctuation, such as ending with a full-stop if they end a sentence.

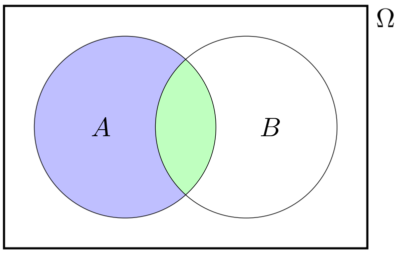
- An extremely rough rule of thumb – which is often, but not always, applicable – is that a good solution should be at least 50% writing and at most 50% equations.
- It's usually best to avoid abbreviations and shorthand, like "LHS" ("left-hand side"), "WLOG" ("without loss of generality"), "iff" ("if and only if"), etc, although specific technical abbreviations from the course, like "PMF" ("probability mass function") are OK. Mathematical shorthand like $\therefore$ or $\Rightarrow$ should be avoided as "word alternatives" ("So", "Thus", "Therefore", "Hence", etc, are better), but are OK in certain circumstances where they have specific technical meanings.
- It's fine – and often good! – to use diagrams to aid your explanation. But a diagram is rarely sufficient by itself without some writing to explain what it shows.

Examples

Here are two solutions for Problem Sheet 3 Question B1. This first solution is a typical sort of solution I might see from a student. The second solution has exactly the same mathematical content, but is more clearly explained. Which do you think is better?

Problem Sheet 3, Question B1. *Suppose A and B are independent events. Show that A and B^c are also independent events.*

Solution 1.



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

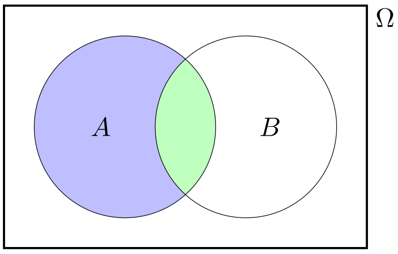
$$\Rightarrow \qquad \begin{aligned} \mathbb{P}(A \cap B^\mathsf{c}) &= \mathbb{P}(A) - \mathbb{P}(A \cap B) \\ &= \mathbb{P}(A) - \mathbb{P}(A)\,\mathbb{P}(B) \\ &= \mathbb{P}(A)(1 - \mathbb{P}(B)) \\ &= \mathbb{P}(A)\,\mathbb{P}(B^\mathsf{c}) \end{aligned}$$

Solution 2. We need to show that A and B^c are independent, which means showing that

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

By splitting the event A up into "the bit in B" and the "the bit not in B", as shown in the Venn diagram below, we have a disjoint union

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



Applying Axiom 3 to this disjoint union, we have

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

Hence, the left-hand side of (∗) is

$$\begin{aligned} \mathbb{P}(A \cap B^{\mathsf{c}}) &= \mathbb{P}(A) - \mathbb{P}(A \cap B) \\ &= \mathbb{P}(A) - \mathbb{P}(A)\,\mathbb{P}(B) \\ &= \mathbb{P}(A)(1 - \mathbb{P}(B)) \\ &= \mathbb{P}(A)\,\mathbb{P}(B^{\mathsf{c}}). \end{aligned}$$

In the second line, we used the fact that A and B are independent, as given in the question, to replace $\mathbb{P}(A \cap B)$ by $\mathbb{P}(A)\,\mathbb{P}(B)$. In the final line, we used the complement rule $\mathbb{P}(B^{\mathsf{c}}) = 1 - \mathbb{P}(B)$. But this is exactly the right-hand side of (∗). Hence, we've shown the left- and right-hand sides of (∗) are equal, and we are done.

Here's another example – Problem Sheet 3, Question B3(b):

Problem Sheet 3, Question B3(b). *Soldiers are asked about their use of illegal drugs, using a so-called "randomised survey". 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?'"*

Suppose that 40% of soldiers respond "Yes". What is the likely proportion of soldiers who have taken illegal drugs in the past 12 months?

Solution 1.

$$\begin{aligned}
&& \mathbb{P}(\text{Yes}) &= \mathbb{P}(C_1)\,\mathbb{P}(\text{Yes} \mid C_1) + \mathbb{P}(C_2)\,\mathbb{P}(\text{Yes} \mid C_2) + \mathbb{P}(C_3)\,\mathbb{P}(\text{Yes} \mid C_3) \\
\Rightarrow && 0.4 &= \tfrac{1}{3} \times 1 + \tfrac{1}{3} \times 0 + \tfrac{1}{3}\,\mathbb{P}(\text{Drugs}) \\
&& &= \tfrac{1}{3} + \tfrac{1}{3}\,\mathbb{P}(\text{Drugs}) \\
\Rightarrow && \mathbb{P}(\text{Drugs}) &= \frac{0.4 - \frac{1}{3}}{\frac{1}{3}} = 20\%
\end{aligned}$$

Solution 2. Let C_1, C_2, C_3 be the events that a soldier picks cards 1, 2, or 3 respectively. These have each probability $\mathbb{P}(C_1) = \mathbb{P}(C_2) = \mathbb{P}(C_3) = \frac{1}{3}$, because the three cards are equally likely.

Let Y be the event that the soldier answers “Yes”. We are told that $\mathbb{P}(Y) = 0.4$.

We know that $\mathbb{P}(Y \mid C_1) = 1$, because a soldier must answer “Yes” to Card 1, and $\mathbb{P}(Y \mid C_2) = 0$, because a soldier must answer “No” to Card 2. For Card 3, a soldier answers “Yes” if they have taken drugs and “No” if they have not, so $\mathbb{P}(Y \mid C_3) = \mathbb{P}(D)$, where $\mathbb{P}(D)$, which we want to find, is the proportion of soldiers who have taken illegal drugs in the past 12 months.

Since C_1, C_2, C_3 make up a partition – a soldier must pick exactly one card – 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 gathered above, we have

$$0.4 = \tfrac{1}{3} \times 1 + \tfrac{1}{3} \times 0 + \tfrac{1}{3}\,\mathbb{P}(D) = \tfrac{1}{3} + \tfrac{1}{3}\,\mathbb{P}(D).$$

Solving this gives

$$\mathbb{P}(D) = \frac{0.4 - \frac{1}{3}}{\frac{1}{3}} = \frac{1}{5} = 20\%.$$

Solutions

This page has the solutions to all the non-assessed questions on Problem Sheet 1 to 5. Solutions are added 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.

There are many ways you get feedback on this module, both group feedback (feedback that is generally relevant to many people) and individual feedback (feedback based specifically on your own approach to the work).

- You will have received both individual and group spoken feedback in your tutorial (the more you speak up in your tutorial, the more individualised the feedback you get in return).
- These solutions include group written feedback on common issues for the class.
- Most importantly, when your work on assessed questions is marked, individual written feedback will be given via the Gradescope site. It is very important that you read that feedback.
- Finally, students who would like even more feedback can discuss their work with me in the “office hours” drop-in sessions.

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.

Group feedback: Many answer for (d) are fine provided you give a justification.

A2. Consider the following data sets of the age of elected politicians on a local council. (The "18–30" bin, for example, means from one's 18th birthday to the moment before one's 30th birthday, so lasts 12 years.)

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		

Age (years)	Frequency	Relative frequency	Frequency density
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
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) What is the modal age bin?

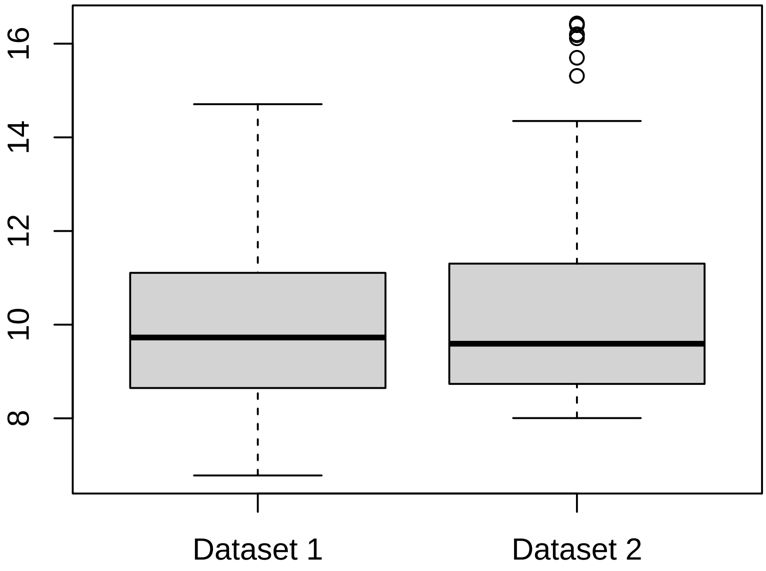
Solution. The bin with the largest frequency density is 45–50, which is therefore the modal bin.

(d) Calculate (the standard approximation of) the mean age of the politicians.

Solution. Pretending that each person is in the centre of their bin, we have

$$\bar{x} = \frac{1}{20}(1 \times 24 + 3 \times 35 + \cdots + 3 \times 65) = \frac{946.5}{20} = 47.3.$$

A3. Consider the two datasets illustrated by the boxplots below. Write down some differences between the two datasets.



Solution. Some answers could be:

- The median and inter-quartile range of Dataset 2 appear to be very slightly larger than those in Dataset 1, although the differences are very small and might not be important in real life.
- Dataset 2 has a few outliers; Dataset 1 has none.
- While Dataset 1 is fairly "balanced" either side of the median, Dataset 2 shows what statisticians call a "positive skew": the data above the median is much more spread out than the data below the median.

Group feedback: You can probably think of other answers.

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

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" above the median or "low outliers" below the median). So the median is robust.

(c) The mean can change a lot if outliers are introduced, especially if the outlier is enormously far our from the data. 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: I find it helpful to suppose I was studying the net worth of people in my tutorial group, and calculating summary statistics. How would those statistics changed change if Elon Musk (founder of Tesla, net worth roughly $200 billion) joined my tutorial group?

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

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

(a) By making a clever choice of (a_i) and (b_i) in the Cauchy–Schwarz inequality, show that $s_{xy}^2 \leq s_x^2 s_y^2$.

Solutions. 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 $s_{xy}^2 \leq s_x^2 s_y^2$.

(b) Hence, show that the correlation r_{xy} satisfies $-1 \leq r_{xy} \leq 1$.

Solutions. Recall the formula for the correlation is

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

We can make part (a) 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: In part (b) there's a temptation to "square-root both sides of the inequality". But you have to be very careful when you do this – make sure you are properly accounting for the positive and negative square roots on both side (if necessary), and where that does or doesn't require reversing the inequality. I recommend leaving the square-root operation until the last possible moment of the proof or, perhaps even better, reasoning through words as I did above.

Remember that you can still attempt part (b) even if you got stuck on part (a).

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 easy 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 4 times.

(a) What would you suggest for a sample space Ω **(i)** if you only care about the total number of heads; **(ii)** if you care about the result of each coin toss?

(b) For each of the cases in part (a), what is $|\Omega|$?

Solution.

(i) We can take $\Omega = \{0, 1, 2, 3, 4\}$, with $|\Omega| = 5$.

(ii) Here, $\Omega = \{\mathrm{HHHH}, \mathrm{HHHT}, \mathrm{HHTH}, \dots, \mathrm{TTTT}\}$ should be the set of all sequences of four "H"s or "T"s. So here, $|\Omega| = 2^4 = 16$.

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. 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) 8^4

Solution.

$$8^4 = 8 \times 8 \times 8 \times 8 = 4096$$

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

Solution.

$$8^{\underline{4}} = 8 \times 7 \times 6 \times 5 = 1680$$

(d) $\dbinom{10}{4}$

Solution.

$$\binom{10}{4} = \frac{10 \times 9 \times 8 \times 7}{4 \times 3 \times 2 \times 1} = 210$$

A4. An urn contains 4 red balls and 6 blue balls. Two balls are drawn from the urn. What is the probability that both balls are red, if the balls are drawn **(a)** with replacement; **(b)** without replacement?

Solution.

(a) There are $|\Omega| = 10^2 = 100$ ways to draw two balls with replacement. There are $|A| = 4^2 = 16$ ways to draw two blue balls. So $\mathbb{P}(A) = \frac{16}{100} = 0.16$.

(b) There are $|\Omega| = 10^{\underline{2}} = 10 \times 9 = 90$ ways to draw two balls without replacement. There are $|A| = 4^{\underline{2}} = 4 \times 3 = 12$ to draw two blue balls. So $\mathbb{P}(A) = \frac{12}{90} = \frac{2}{15} = 0.133$.

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

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

Solution. 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 $A = \emptyset$. Since $\emptyset^{\mathsf{c}} = \Omega$ and Axiom 2 tells us that $\mathbb{P}(\Omega) = 1$, the result follows.

Group feedback: With this, and most "prove from the axioms" questions, the key is to find a relevant disjoint union, which then allows us to use Axiom 3. So if we can find $C = A \cup B$ as a disjoint union (hopefully containing some events relevant to the question at hand), Axiom 3 allows us to write $\mathbb{P}(C) = \mathbb{P}(A) + \mathbb{P}(B)$.

(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. 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

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

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 $B = D \cup E$ 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

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

as required.

(b) Using your result from part (a), the two-event addition rule, the distributive law, and the two-event addition rule again, 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 $\mathbb{P}(C \cap (D \cup E))$. 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

$\mathbb{P}((C \cap D) \cup (C \cap E))$. But this is another union of two events again, this time with $A = C \cap D$ and $B = C \cap E$. So the two-event addition rule gives

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

since $(C \cap D) \cap (C \cap E) = C \cap D \cap E$.

Finally, we put this all together, and get

$$\begin{aligned}&\mathbb{P}(C \cup D \cup E)\\ &\quad = \mathbb{P}(C) + (\mathbb{P}(D) + \mathbb{P}(E) - \mathbb{P}(D \cap E)) - (\mathbb{P}(C \cap D) + \mathbb{P}(C \cap E) - \mathbb{P}(C \cap D \cap E))\\ &\quad = \mathbb{P}(C) + \mathbb{P}(D) + \mathbb{P}(E) - \mathbb{P}(C \cap D) - \mathbb{P}(C \cap E) - \mathbb{P}(D \cap E) + \mathbb{P}(C \cap D \cap E),\end{aligned}$$

which is what we wanted.

B3. Suppose we pick a number at random from the set $\{1, 2, \dots, 2022\}$.

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

Solution. The sample space is $\Omega = \{1, 2, \dots, 2022\}$. Clearly $|\Omega| = 2022$. Further, the event in question is $A = \{5, 10, \dots, 2020\}$ of numbers up to 2022 that are divisible by 5. Thus $|A|$ is the largest integer no bigger than $\frac{2022}{5} = 404.4$, which is 404, as this is how many times 5 "goes into" 2022. Hence

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{404}{2022} = 0.1998,$$

just a tiny bit smaller than $\frac{1}{5}$.

Group feedback: With these "classical probability" questions, the steps should always be:

1. State clearly what the sample space Ω is.
2. Count how many outcomes $|\Omega|$ are in the sample space.
3. State clearly what the event A is.
4. Count how many outcomes $|A|$ are in the event.
5. The desired probability is then $\mathbb{P}(A) = |A|/|\Omega|$.

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

Solution. With the same Ω and A, now let B be the numbers up to 2022 divisible by 7; so we're looking for $\mathbb{P}(A \cup B)$. By the addition rule for unions, this is

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

We already know $\mathbb{P}(A) = \frac{404}{2022}$, so need to find out $\mathbb{P}(B)$ and $\mathbb{P}(A \cap B)$.

As before, $|B|$ is the largest integer no bigger that $\frac{2022}{7} = 288.9$, which is 288. So So $\mathbb{P}(B) = \frac{288}{2022}$. Now, $A \cap B$ 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 $\frac{2022}{35} = 57.8$ rounded down, so $\mathbb{P}(A \cap B) = \frac{57}{2022}$.

So finally, we have

$$\mathbb{P}(A \cup B) = \frac{404}{2022} + \frac{288}{2022} - \frac{57}{2022} = \frac{635}{2022} = 0.314.$$

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 sample space 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

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{8 \times 2 \times 1 \times 5 \times 4 \times 3 \times 2 \times 1}{8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1} = \frac{2 \times 1}{7 \times 6} = \frac{1}{21}.$$

Group feedback: As we have discussed recently, often "classical probability" problems can be solved by the step-by-step "chain rule" method. Can you use a chain rule argument to find the same answer as

$$\mathbb{P}(A) = 1 \times \frac{2}{7} \times \frac{1}{6} \times 1 \times 1 \times 1 \times 1 \times 1 = \frac{1}{21}?$$

(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 Brook's and Chris's 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, leaving $6 - 3 = 3$ choices.
  - If Brook 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, leaving $6 - 4 = 2$ choices.
- 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, in a previous tutorial, 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, then Brook will have 6 gaps (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 assuming Brook was not next to Ashley). 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. A “random digit” is a number chosen at random from $\{0, 1, \dots, 9\}$, each with equal probability. A statistician chooses n random digits (with replacement).

(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?

Solution. The sample space is $\Omega = \{0, 1, \dots, 9\}^n$, the set of length-$n$ sequences of digits between 0 and 9. The number of these is $|\Omega| = 10^n$, as there are 10 choices for each of the n digits.

The event A_k is $\{0, 1, \dots, k\}^n$, the set of length-n sequences of digits that are between 0 and k. The number of these is $|A_k| = (k + 1)^n$. (Note that it’s $k + 1$ because we’re allowing 0 as well.)

Hence, the probability is

$$\mathbb{P}(A_k) = \frac{|A_k|}{|\Omega|} = \frac{(k + 1)^n}{10^n}.$$

(b) Let B_k be the event that the largest digit chosen is equal to k. By finding a relationship between B_k, A_{k-1} and A_k, or otherwise, show that

$$\mathbb{P}(B_k) = \frac{(k + 1)^n - k^n}{10^n}.$$

Solution. Consider the event A_k that all the digits are at most k. Within A_k, *either* one or more of the digits is k, in which case that is the largest digit we are in B_k; *or* none of the digits are k, in which case they are all at most $k - 1$, and we are in A_{k-1}, *but not both.* Hence we have a disjoint union

$$A_k = B_k \cup A_{k-1}.$$

Applying Axiom 3 gives

$$\mathbb{P}(A_k) = \mathbb{P}(B_k) + \mathbb{P}(A_{k-1}).$$

Rearranging this gives

$$\mathbb{P}(B_k) = \mathbb{P}(A_k) - \mathbb{P}(A_{k-1}).$$

Substituting in the answer from part (a) gives

$$\mathbb{P}(B_k) = \frac{(k+1)^n}{10^n} - \frac{(k-1+1)^n}{10^n} = \frac{(k+1)^n - k^n}{10^n}.$$

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

$$\begin{aligned}\mathbb{P}(\text{first Heart}) &= \frac{13}{52} = \frac{1}{4}\\ \mathbb{P}(\text{second Ace}) &= \frac{4}{52} = \frac{1}{13}\\ \mathbb{P}(\text{first Heart and second Ace}) &= \frac{12\times 4 + 1\times 3}{52\times 51} = \frac{51}{52\times 51} = \frac{1}{52}\end{aligned}$$

Here, the 12×4 counted "a non-Ace Heart, followed by an Ace", while the 1×3 counted "the Ace of Hearts, followed by a non-Heart Ace". 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. Consider rolling two dice. 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. You may assume the dice rolls are independent; so, in particular, events A and B are independent.

(a) Are A and C independent?

Solution. Let us first note that $\mathbb{P}(A) = \mathbb{P}(B) = \frac{3}{6} = \frac{1}{2}$. It's also the case that $\mathbb{P}(C) = \frac{18}{36} = \frac{1}{2}$, for example by counting the 18 even outcomes out of the 36 equally likely possibilities.

We need to test is $\mathbb{P}(A \cap C) = \mathbb{P}(A)\,\mathbb{P}(C) = \frac{1}{4}$ or not. By counting from the 36 possibilities, we see that indeed $\mathbb{P}(A \cap C) = \frac{9}{36} = \frac{1}{4}$. Alternatively, we could note that $\mathbb{P}(C \mid A) = \mathbb{P}(B) = \frac{1}{2}$, since if the first dice is even, the second must be also even to get an even total. Then $\mathbb{P}(A \cap C) = \mathbb{P}(A)\,\mathbb{P}(C \mid A) = \frac{1}{4}$.

So the events are independent.

(b) Are B and C independent?

Solution. Yes. The solution is essentially identical to part (a).

(c) Is it true that $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\,\mathbb{P}(B)\,\mathbb{P}(C)$?

Solution. By checking the 36 possibilities, one sees that

$$\mathbb{P}(A \cap B \cap C) = \frac{9}{36} = \frac{1}{4} \neq \frac{1}{8} = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \mathbb{P}(A)\,\mathbb{P}(B)\,\mathbb{P}(C).$$

Alternatively, note that the total being even is certain if both dice rolls are certain, so

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A \cap B)\,\mathbb{P}(C \mid A \cap B) = \frac{1}{4} \times 1 = \frac{1}{4},$$

to get the same result.

Group feedback: This shows that just because events are "pairwise independent", it does not mean they are "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,

$$\operatorname{Var}(X) = \mathbb{E}X^2 - \mu^2 = 0.775 - 0.45^2 = 0.5725.$$

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	0.1

This random variable has $\mathbb{E}X = 4.3$. Find the values of a and b.

Solution. First, a PMF must sum to 1, so

$$1 = 0.1 + 0.2 + 0.1 + b + 0.1,$$

so $b = 0.5$.

Second, the expectation is

$$\mathbb{E}X = 1 \times 0.1 + 2 \times 0.2 + 4 \times 0.1 + 5b + 0.1a = 3.6 + 0.1a = 4.3.$$

So $a = 7$.

A5. A 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. The variance of the daily maximum temperature measured in degrees Celsius is 10.4.

(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 when measured in degrees Fahrenheit?

Solution. For the variance,

$$\operatorname{Var}(T_F) = \operatorname{Var}\left(\tfrac{9}{5}T_C + 32\right) = \left(\tfrac{9}{5}\right)^2 \operatorname{Var}(T_C) = \tfrac{81}{25}\operatorname{Var}(T_C).$$

So the answer is $\frac{81}{25} \times 10.4 = 33.7$.

B1. Suppose A and B are independent events. Show that A and B^{c} are also independent events.

Solution. We know that

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

because A and B are independent. We need to show that A and B^{c} are independent, which means showing that

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

Note that

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

and the union is disjoint, so by Axiom 3,

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

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, crucially we used the fact that A and B are independent to replace $\mathbb{P}(A \cap B)$ by $\mathbb{P}(A)\,\mathbb{P}(B)$.

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

Hence, we've shown the left- and right-hand sides of $(*)$ are equal, and we are done.

B2. You are dealt a hand of 13 cards from a 52-card deck. Let $E_{\mathrm{A}}, E_{\mathrm{K}}, E_{\mathrm{Q}}, E_{\mathrm{J}}$ respectively be the events that your hand contains the Ace, King, Queen and Jack of Spades.

(a) What is $\mathbb{P}(E_{\mathrm{A}})$, the probability that your hand contains the Ace of Spades?

Solution. There are 52 cards of which 13 will end up in my hand, so $\mathbb{P}(E_{\mathrm{A}}) = \frac{13}{52}$.

(b) Explain why $\mathbb{P}(E_{\mathrm{K}} \mid E_{\mathrm{A}}) = \frac{12}{51}$.

Solution. Given I have the Ace of Spades, there are $52 - 1 = 51$ cards left available, of which $13 - 1 = 12$ will end up in my hand, so $\mathbb{P}(E_{\mathrm{K}} \mid E_{\mathrm{A}}) = \frac{12}{51}$.

(c) Using the chain rule, calculate the probability that your hand contains all four of the Ace, King, Queen and Jack of Spades.

Solution. Continuing the logic of part (b), we have

$$\mathbb{P}(E_{\mathrm{Q}} \mid E_{\mathrm{A}} \cap E_{\mathrm{K}}) = \frac{11}{50} \qquad \mathbb{P}(E_{\mathrm{J}} \mid E_{\mathrm{A}} \cap E_{\mathrm{K}} \cap E_{\mathrm{Q}}) = \frac{10}{49}.$$

Using the chain rule,

$$\begin{aligned} P(E_{\mathrm{A}} \cap E_{\mathrm{K}} \cap E_{\mathrm{Q}} \cap E_{\mathrm{J}}) &= \mathbb{P}(E_{\mathrm{A}})\,\mathbb{P}(E_{\mathrm{K}} \mid E_{\mathrm{A}})\,\mathbb{P}(E_{\mathrm{Q}} \mid E_{\mathrm{A}} \cap E_{\mathrm{K}})\,\mathbb{P}(E_{\mathrm{J}} \mid E_{\mathrm{A}} \cap E_{\mathrm{K}} \cap E_{\mathrm{Q}}) \\ &= \frac{13}{52} \times \frac{12}{51} \times \frac{11}{50} \times \frac{10}{49} = \frac{13 \times 12 \times 11 \times 10}{52 \times 51 \times 50 \times 49}, \end{aligned}$$

which is the same as we got in lectures.

(d) Check that your answer agrees with the answer we found by classical probability methods in Example 6.4 in Lecture 6. Which method do you prefer?

Solution. Personally, I slightly prefer this answer – it seems more obvious how the answer relates to the method, whereas in lectures a lot of terms in a ratio of binomial coefficients "magically" cancelled out. Your mileage may vary.

B3. Soldiers are asked about their use of illegal drugs, using a so-called "randomised survey". 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) What are some advantages or disadvantages of performing the experiment this way?

Solution. The main advantage is that it seems likely that a soldier might want to lie in answer to a "straight question", given that if their superiors discovered they had taken illegal drugs, there could be very serious consequences. This method allows a certain “plausible deniability”: just because the soldier answers “Yes”, we cannot know for sure whether they have taken illegal drugs or merely picked the “Yes” card. Thus we might hope to get more honest answers this way. Perhaps you can think of other advantages.

There could be disadvantages. The complicated set-up of the experiment could lead to the subjects (or experimenters) making an error. The scientists good be “lulled into a false sense of security” of thinking they get fully honest answers, when soldiers picking card 3 might still choose to lie. Perhaps you can think of other disadvantages.

(b) Suppose that 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 $p = \frac{1}{5} = 20\%$.

(c) 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

$$\mathbb{P}(D \mid Y) = \frac{0.2 \times \frac{2}{3}}{0.4} = \frac{1}{3}.$$

B4. A random variable X_n is said to follow the *discrete uniform distribution* on $\{1, 2, \dots, n\}$ if each of the n values in that set $\{1, 2, \dots, n\}$ is equally likely.

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

Solution. We have $p(x) = \frac{1}{n}$ for $x = 1, 2, \dots, n$. So the expectation is

$$\mathbb{E}X = \sum_{x=1}^{n} x\,\frac{1}{n} = \frac{1}{n}\sum_{x=1}^{n} x = \frac{1}{n}\,\frac{n(n+1)}{2} = \frac{n+1}{2}.$$

(b) Find the variance of X_n.

Solution. It turns out to be much easier to use the computational formual $\mathrm{Var}(X) = \mathbb{E}X^2 - \mu^2$. First,

$$\mathbb{E}X^2 = \sum_{x=1}^{n} x^2 \, \frac{1}{n} = \frac{1}{n} \sum_{x=1}^{n} x^2 = \frac{1}{n} \, \frac{n(n+1)(2n+1)}{6} = \frac{(n+1)(2n+1)}{6}.$$

Then using $\mu = (n+1)/2$ from part (a), we have

$$\mathrm{Var}(X) = \frac{(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2 = \frac{(n+1)(4n+2-3n-3)}{12} = \frac{(n+1)(n-1)}{12} = \frac{n^2-1}{12}$$

(c) Let Y be a discrete uniform distribution on $b-a+1$ values $\{a, a+1, a+2, \dots, b-1, b\}$, 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.

*[**Note:** "$b-a+1$ values" is correct, but this was wrong earlier.]*

Solution. If we take $n = b-a+1$, then Y has the same distribution as $X_n + (a-1)$. This is because $x = 1$ maps to $y = 1+(a-1) = a$; $x = 2$ maps to $y = 2+(a-1) = a+1$; and so on; up to $x = n$ mapping to $y = n+(a-1) = (b-a+1)+(a-1) = b$. So the ranges match up perfectly.

Thus we have

$$\mathbb{E}Y = \mathbb{E}(X_{b-a+1} + (a-1)) = \mathbb{E}X_{b-a+1} + (a-1) = \frac{b-a+1+1}{2} + (a-1) = \frac{a+b}{2}$$

and

$$\mathrm{Var}(Y) = \mathrm{Var}\left(X_{b-a+1} + (a-1)\right) = \mathrm{Var}(X_{b-a+1}) = \frac{(b-a+1)^2-1}{12}.$$

You can rearrange the variance a bit if you like, but it doesn't really get any nicer.

You may use without proof the standard results

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

B5. A gambling game works as follows. You keep tossing a fair coin until you first get a Head. If the first Head comes up 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 observation that, despite the expected winnings from this game being infinite, few people would be prepared to play this game for, say, £100, and almost no one 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.

Problem Sheet 4

A1. Let $X \sim \mathrm{Bin}(20, 0.4)$. Calculate

(a) $\mathbb{P}(X = 8)$

Solution.

$$\mathbb{P}(X = 8) = \binom{20}{8} 0.4^8 \times 0.6^{12} = 0.180.$$

(b) $\mathbb{P}(8 \leq X \leq 11)$

Solution.

$$\begin{aligned}\mathbb{P}(8 \leq X \leq 11) &= \mathbb{P}(X = 8) + \mathbb{P}(X = 9) + \mathbb{P}(X = 10) + \mathbb{P}(X = 11) \\ &= \binom{20}{8} 0.4^8 \times 0.6^{12} + \binom{20}{9} 0.4^9 \times 0.6^{11} + \binom{20}{10} 0.4^10 \times 0.6^{10} + \binom{20}{11} 0.4^8 \times 0.6^{11} \\ &= 0.180 + 0.160 + 0.117 + 0.071 \\ &= 0.528.\end{aligned}$$

(c) $\mathbb{E}X$

Solution. $\mathbb{E}X = 20 \times 0.4 = 8$.

A2. Let $X \sim \mathrm{Geom}(0.2)$. Calculate

(a) $\mathbb{P}(X = 2)$

Solution. $\mathbb{P}(X = 2) = 0.8^1 \times 0.2^1 = 0.16$.

(b) $\mathbb{P}(X \geq 3)$

Solution. $\mathbb{P}(X \geq 3) = 1 - \mathbb{P}(X = 1) - \mathbb{P}(X = 2) = 1 - 0.2 - 0.8 \times 0.2 = 0.64$.

(c) $\mathrm{Var}(X)$

Solution. $\mathrm{Var}(X) = \dfrac{1 - 0.2}{0.2^2} = 20$.

A3. Let $X \sim \mathrm{Po}(2.5)$. Calculate

(a) $\mathbb{P}(X = 3)$

Solution. $\mathbb{P}(X = 3) = \mathrm{e}^{-2.5} \dfrac{2.5^3}{3!} = 0.214$.

(b) $\mathbb{P}(X \geq \mathbb{E}X)$

Solution. First, $\mathbb{E}X = 2.5$. So

$$\begin{aligned}\mathbb{P}(X \geq \mathbb{E}X) &= \mathbb{P}(X \geq 2.5) \\ &= 1 - \mathbb{P}(X = 0) - \mathbb{P}(X = 1) - \mathbb{P}(X = 2) \\ &= 1 - \mathrm{e}^{-2.5} - 2.5\mathrm{e}^{-2.5} - \frac{2.5^2}{2}\mathrm{e}^{-2.5} \\ &= 1 - 0.082 - 0.204 - 0.257 \\ &= 0.456.\end{aligned}$$

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.

Solution. The total of the joint PMF is

$$2k+2k+k+k+3k+k+k+k+k+2k=15k$$

which must be 1, so $k=\frac{1}{15}$.

(b) Find the marginal PMFs of X and Y.

Solution. By summing across the rows and down the columns, respectively, we get this:

$p_{X,Y}(x,y)$	$y=0$	$y=1$	$y=2$	$y=3$	$p_X(x)$
$x=0$	$\frac{2}{15}$	$\frac{2}{15}$	$\frac{1}{15}$	0	$\frac{5}{15}$
$x=1$	$\frac{1}{15}$	$\frac{3}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{6}{15}$
$x=2$	0	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{4}{15}$
$p_Y(y)$	$\frac{3}{15}$	$\frac{6}{15}$	$\frac{3}{15}$	$\frac{3}{15}$	

(c) What is the conditional distribution of Y given $X=1$?

Solution. We get this by taking the $x=1$ row of the table, than normalising it by dividing through by $p_X(1)=\frac{6}{15}$. This gives

$$p_{Y\mid X}(0\mid 1)=\frac{1}{6} \qquad p_{Y\mid X}(1\mid 1)=\frac{3}{6} \qquad p_{Y\mid X}(2\mid 1)=\frac{1}{6} \qquad p_{Y\mid X}(3\mid 1)=\frac{1}{6}.$$

(d) Are X and Y independent?

Solution. No. For one example, $p_{X,Y}(0,0)=\frac{2}{15}$, while $p_X(0)\,p_Y(0)=\frac{5}{15}\times\frac{3}{15}=\frac{1}{15}$, so they are not equal.

Questions B3, B4 and B5 on Problem Sheet 4 were discussed in the online tutorial. A video recording of the tutorial is available on Minerva.

B1. Calculate the CDF $F(x)=\mathbb{P}(X\leq x)$ of the geometric distribution…

(a) …by summing the PMF;

Solution. We have,using the standard formula for the sum of a finite geometric progression,

$$\begin{aligned}
F(x) &= \sum_{y=1}^{x} p(y) \\
&= \sum_{y=1}^{x} (1-p)^{y-1} p \\
&= \frac{p(1-(1-p)^x)}{1-(1-p)} \\
&= \frac{p(1-(1-p)^x)}{p} \\
&= 1-(1-p)^x .
\end{aligned}$$

(b) ...by explaining how the "number of trials until success" definition tells us what $1 - F(x) = \mathbb{P}(X > x)$ must be.

Solution. Note that $1 - F(x) = \mathbb{P}(X > x)$ is precisely the probability that the first x trials are failures, and hence that the first success comes strictly after the xth trial. The probability that the first x trials are failures is $(1-p)^x$. So $F(x) = 1 - (1-p)^x$.

(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?

Solution. Let $X \sim \text{Geom}(\frac{1}{36})$. Then

$$\begin{aligned}
\mathbb{P}(20 \leq X \leq 40) &= \mathbb{P}(X \leq 40) - \mathbb{P}(X \leq 19) \\
&= F(40) - F(19) \\
&= \left(1 - (1-\tfrac{1}{36})^{40}\right) - \left(1 - (1-\tfrac{1}{36})^{19}\right) \\
&= 0.676 - 0.414 \\
&= 0.261.
\end{aligned}$$

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.

Solution. Having $Y = y$ requires y consecutive failures immediately followed by a success. So $p_Y(y) = (1-p)^y p$.

(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 $\text{Var}(X) = (1-p)/p^2$.

Solution. If $X \sim \text{Geom}(p)$ under the standard definition, then (as we saw in the notes) Y has the same distribution as $X - 1$. Therefore,

$$\mathbb{E}Y = \mathbb{E}(X-1) = \mathbb{E}X - 1 = \frac{1}{p} - 1 = \frac{1-p}{p}$$

and

$$\operatorname{Var}(Y) = \operatorname{Var}(X-1) = \operatorname{Var}(X) = \frac{1-p}{p^2}.$$

B3 Let $X \sim \operatorname{Po}(\lambda)$.

(a) Show that $\mathbb{E}X(X-1) = \lambda^2$. You may use the Taylor series for the exponential,

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

Solution. We follow exactly the method used to calculate $\mathbb{E}X$ in the notes. We have

$$\begin{aligned}
\mathbb{E}X(X-1) &= \sum_{x=0}^{\infty} x(x-1)\,\mathrm{e}^{-\lambda}\frac{\lambda^x}{x!} \\
&= \lambda^2 \mathrm{e}^{-\lambda} \sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} \\
&= \lambda^2 \mathrm{e}^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} \\
&= \lambda^2 \mathrm{e}^{-\lambda}\,\mathrm{e}^{\lambda} \\
&= \lambda^2.
\end{aligned}$$

In the second line, we took a λ^2 and a $\mathrm{e}^{-\lambda}$ outside the brackets; cancelled the x and $x-1$ out of the $x!$; and removed the $x=0$ and $x=1$ terms from the sum, since they were 0 anyway. In the third line, we re-indexed the sum by setting $y = x-2$. In the fourth line, we used the Taylor series for the exponential

(b) Hence show that $\operatorname{Var}(X) = \lambda$. You may use the fact, proved in the notes, that $\mathbb{E}X = \lambda$.

Solution. We know from part (a) that

$$\mathbb{E}X(X-1) = \mathbb{E}(X^2 - X) = \mathbb{E}X^2 - \mathbb{E}X = \mathbb{E}X^2 - \lambda = \lambda^2,$$

which gives $\mathbb{E}X^2 = \lambda^2 + \lambda$. We can then use the computational formula for the variance to get

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

B4. Each week in the UK about 15 million Lotto tickets are sold. As we saw in Lecture 6, 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.

Solution. We assume that each ticket is uniformly randomly chosen from all possible tickets, independent of all other tickets. Then the number of winners is $X \sim \operatorname{Bin}(15 \text{ million}, 1/(45 \text{ million}))$. It will be convenient to use a Poisson approximation with rate

$$\lambda = 15 \text{ million} \times \frac{1}{45 \text{ million}} = \tfrac{1}{3}.$$

The probability there is a roll-over is

$$\mathbb{P}(X = 0) \approx \mathrm{e}^{-1/3} = 0.72.$$

The probability there is a unique jackpot winner is

$$\mathbb{P}(X = 1) \approx \tfrac{1}{3}\mathrm{e}^{-1/3} = 0.24.$$

The probability there are multiple winners is

$$\mathbb{P}(X \geq 2) = 1 - \mathbb{P}(X = 0) - \mathbb{P}(X = 1) = 0.04.$$

B5. 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.

Solution. For all these questions, we need to fill in a table for the joint PMF, where the columns sum to $p_X(0) = p_X(1) = \frac{1}{2}$ and the rows sum to $p_Y(0) = p_Y(1) = \frac{1}{2}$.

$p_{X,Y}(x,y)$	$x=0$	$x=1$	$p_Y(y)$
$y=0$			$\frac{1}{2}$
$y=1$			$\frac{1}{2}$
$p_Y(y)$	$\frac{1}{2}$	$\frac{1}{2}$	

If X and Y are independent, we have $p_{X,Y}(x,y) = p_X(x)\,p_Y(y)$; so, for example, $p_{X,Y}(0,0) = p_X(0)\,p_Y(0) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$. In fact, all the entries in the joint PMF table are $\frac{1}{4}$.

$p_{X,Y}(x,y)$	$x=0$	$x=1$	$p_Y(y)$
$y=0$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
$y=1$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
$p_Y(y)$	$\frac{1}{2}$	$\frac{1}{2}$	

(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.

Solution. We still need the rows and columns to add up to $\frac{1}{2}$, but we want low values of X (that is, 0) to be more likely to occur alongside low values of Y (that is, 0), and high values of X (that is, 1) alongside high values of Y (that is, 1). One was to do this is

$p_{X,Y}(x,y)$	$x=0$	$x=1$	$p_Y(y)$
$y=0$	$\frac{1}{2}$	0	$\frac{1}{2}$
$y=1$	0	$\frac{1}{2}$	$\frac{1}{2}$

$p_{X,Y}(x,y)$	$x=0$	$x=1$	$p_Y(y)$
$p_Y(y)$	$\frac{1}{2}$	$\frac{1}{2}$	

A single table like that is a perfectly sufficient answer. But, in fact, any table of the form

$p_{X,Y}(x,y)$	$x=0$	$x=1$	$p_Y(y)$
$y=0$	a	$\frac{1}{2}-a$	$\frac{1}{2}$
$y=1$	$\frac{1}{2}-a$	a	$\frac{1}{2}$
$p_Y(y)$	$\frac{1}{2}$	$\frac{1}{2}$	

for $\frac{1}{4} < a \leq \frac{1}{2}$ will do. This has

$$\mathbb{E}XY = \sum_{x,y} xy\, p_{X,Y}(x,y) = p_{X,Y}(1,1) = a,$$

as $x = y = 1$ is the only nonzero term in the sum. This means the covariance is, by the computational formula,

$$\mathrm{Cov}(X,Y) = \mathbb{E}XY - \mu_X\mu_Y = a - \tfrac{1}{2} \times \tfrac{1}{2} = a - \tfrac{1}{4}.$$

So the covariance is positive for $a > \frac{1}{4}$, so the correlation is too.

(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.

Solution. For example

$p_{X,Y}(x,y)$	$x=0$	$x=1$	$p_Y(y)$
$y=0$	0	$\frac{1}{2}$	$\frac{1}{2}$
$y=1$	$\frac{1}{2}$	0	$\frac{1}{2}$
$p_Y(y)$	$\frac{1}{2}$	$\frac{1}{2}$	

Alternatively, any table of the form

$p_{X,Y}(x,y)$	$x=0$	$x=1$	$p_Y(y)$
$y=0$	a	$\frac{1}{2}-a$	$\frac{1}{2}$
$y=1$	$\frac{1}{2}-a$	a	$\frac{1}{2}$
$p_Y(y)$	$\frac{1}{2}$	$\frac{1}{2}$	

for $0 \leq a < \frac{1}{4}$ will have negative covariance $a - \frac{1}{4}$, so negative correlation.