MATH1710 Probability and Statistics I

Matthew Aldridge

University of Leeds, 2021–22

Contents

Schedule

MATH1710 begins on Monday 27 September.

About MATH2750

Organisation of MATH2750

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

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

The core teaching team are:

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

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

Notes and videos

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

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

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

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

Problem sheets

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

Problem Sheet	Sections covered	Assessed work due
1	1	Friday 8 October (week 2)
2	2 and 3	Friday 22 October (week 4)
3	4 and 5	Friday 5 November (week 6)
4	6 and 7	Friday 19 November (week 8)
5	8, 9 and 10	Friday 3 December (week 10)

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

Lectures

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

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

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

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

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

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

Tutorials

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

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

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

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

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

R worksheets

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

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

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

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

Office hours

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

There will be two office hours per week: Wednesdays at 1000 and at 1200. (For boring reasons, the 1000 sessions appear on the timetable for MATH2700 students and the 1200 sessions appear on the timetable for MATH1710 students, but I'm happy for anyone to attend either hour.)

Time management

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

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

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

Exam

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

Who should I ask about...?

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

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

Content of MATH2750

Prerequisites

The formal prerequisite for MATH1710 is "Grade B in A-level Mathematics or equivalent". We'll assume you have some basic school-level maths knowledge, but we don't assume you've studied probability or statistics in detail before (although we recognise that many of you will have). If you have studied probability and/or statistics at A-level (or post-16 equivalent) level, you'll recognise

some of the material in this module; however you should find that we go deeper in some areas, and that we treat the material through with a greater deal of mathematical formality and rigour. "Rigour" here means precisely stating our assumptions, and carefully *proving* how other statements follow from those assumptions.

Syllabus

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

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

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

Books

You can do well on this module by reading the notes and watching the videos, attending the lectures and tutorials, and working on the problem sheets and R worksheets, without needing to do any further reading beyond this. However, students can benefit from optional extra background reading or an alternative view on the material, especially in the parts of the module on probability.

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

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

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

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

On Bayesian statistics, I recommend:

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

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

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

About these notes

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

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

Part I: EDA

Chapter 1

Exploratory data analysis

1.1 What is EDA?

Statistics is the study of data. **Exploratory data analysis** (or **EDA**, for short) is the part of statistics concerned with taking a "first look" at some data. Later, toward the end of this course, we will see more detailed and complex ways of building models for data, and in MATH1712 Probability and Statistics 2 (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?*
- *How was the data collected?*
- *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 them. Perhaps it is a genuine-but-unusual datapoint (which is useful for understand the extremes of the data), or perhaps there is an extraordinary explanation (maybe 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.

1.2 What is R?

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

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

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

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

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

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

1.3 Summary statistics and boxplots

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

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

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

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

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

Some measures of centrality are the following.

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

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

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

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

Example 1.1. EXAMPLE

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

Two other common terms: $q(\frac{3}{4})$ is called the **upper quartile** and $q(\frac{1}{4})$ is called the **lower quartile** (note "quartile" – as in "quarter" – not "quantile", here).

Some measures of spread are:

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

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

The **sample variance** is

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

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

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

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

Here, the first line is the definitional formula; the second line is from expanding out the bracket; the third line is taking the sum term-by-term; the fourth line takes any constants (things not involving i) outside the sums; the fifth line uses $\sum_{i=1}^{n} x_i = n\bar{x}$, from the definition of the mean, and $\sum_{i=1}^{n} 1 = 1 + 1 + \cdots 1 = n$; and the sixth line simplifies the final two terms.

This has left us with

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

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

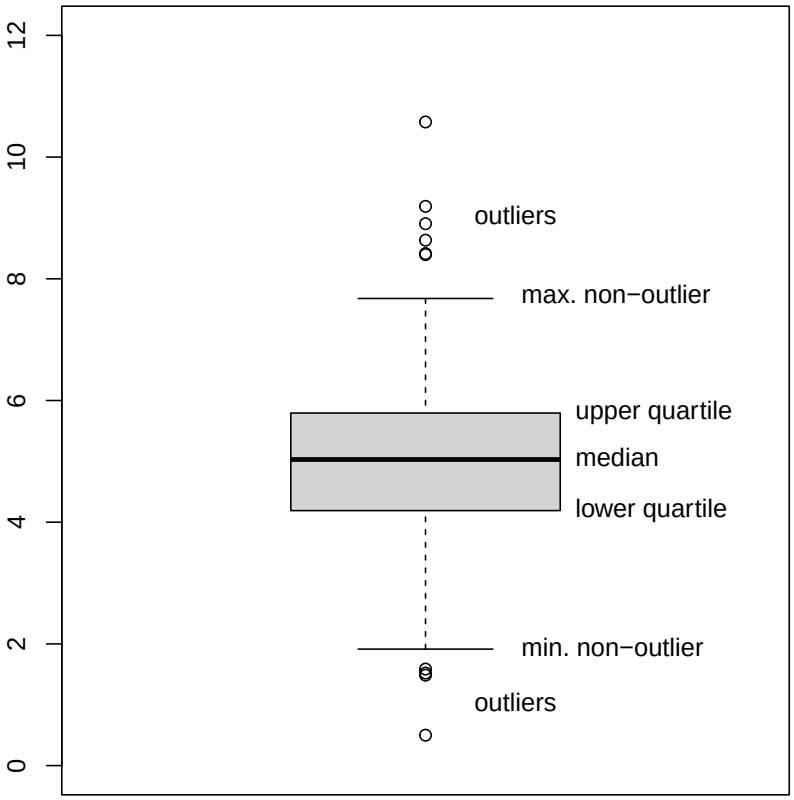
The following R code reads in some data which has the daily average temperature in Leeds, divided into months. We can find, for example, ???????????????? ???????????/

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

A boxplot is drawn as follows:

- The vertical axis represents the data values.
- Draw a box from the lower quartile $q(\frac{1}{4})$ to the median $q(\frac{1}{2})$.
- Draw another box on top of this from the median $q(\frac{1}{2})$ to the upper quartile $q(\frac{3}{4})$. Note that size of these two boxes put together is the interquartile range.

- Decide which datapoints are outliers, and plot these with circles. (The R default is that any data point less than $q(\frac{1}{4}) - 1.5 \times \text{IQR}$ or greater than $q(\frac{3}{4}) + 1.5 \times \text{IQR}$ is an outlier.)
- Out from the two previous boxes, draw "whiskers" to the smallest and largest non-outlier datapoints.



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

```
#boxplot(temp$jul, temp$sep,
#         names = c("July", "September"),
#         ylab = "Temperature (degrees C)")
```

1.4 Binned data and histograms

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

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

Time	Frequency	Relative frequency
0–5 minutes	4	0.04
5–10 minutes	8	0.08
10–15 minutes	21	0.21
15–30 minutes	42	0.42
30–45 minutes	15	0.15
45–60 minutes	8	0.08
60–120 minutes	2	0.02
Total	100	1

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

What is the median journey length? Well, we don't know exactly, but $0.04 + 0.08 + 0.21$ is less than 0.5, while $0.04 + 0.08 + 0.21 + 0.42$ is greater than 0.5. So we know that the median student is in the "10–15 minute" bin, and we can say that the median journey length is between 10 and 15 minutes.

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

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

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

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

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

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

More generally, if m_j is the midpoint of bin j and f_j its frequency, then we can calculate the binned mean and binned variance by

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

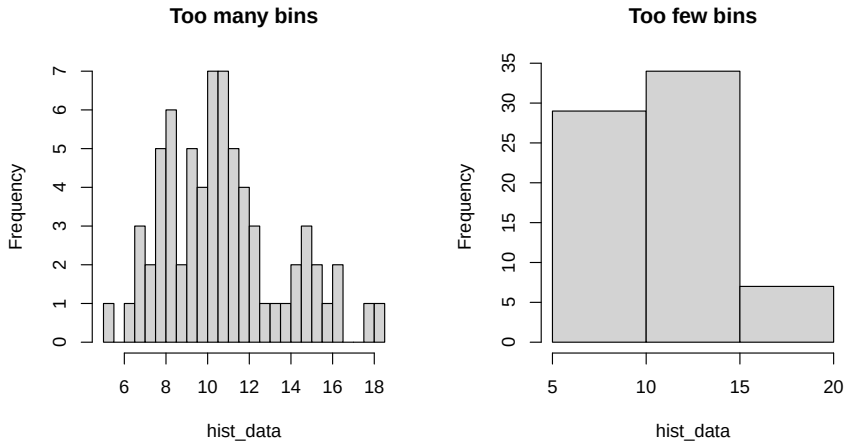
Data in bins can be illustrated with a **histogram**. A histogram has the measurement on the x-axis, with one bar across the width of each bin, with bars drawn up to the height of the corresponding frequency density. Note that this means that the area of the bar is exactly the relative frequency of the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

Here is a histogram for our journey-time data:

PICTURE

Often we draw histograms because the data was collected in bins. But even when we have exact data, we might choose to divide it into bins for the purposes of drawing a histogram. In this case we have to decide where to put the "breaks" between the bins. Too many breaks too close together, and the small number of observations in each bin will give "noisy" results (see left); too few breaks too far apart, and the histogram will lose detail (see right).

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



1.5 Multiple measurements and scatterplots

Often, more than one piece of data is collected from each subject, and we wish to compare that data.

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?

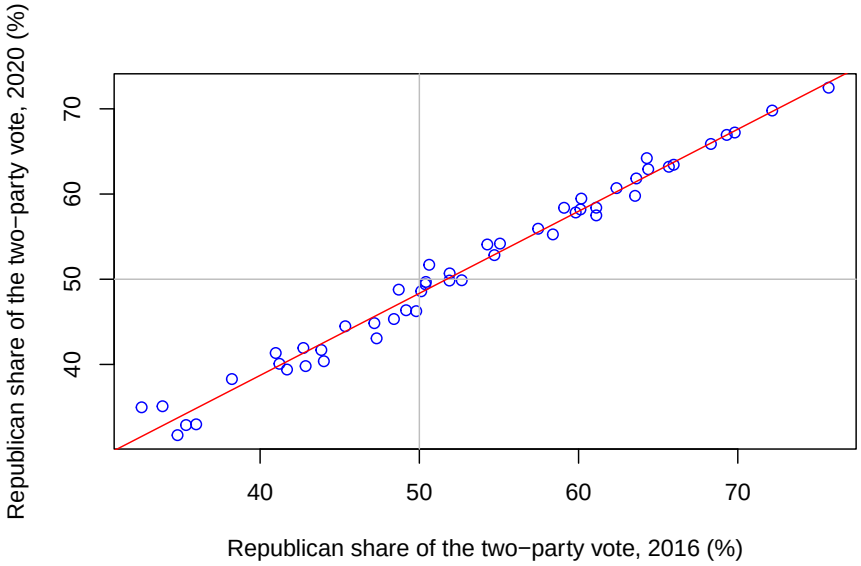
A good way to visualise the relationship between two variables is to use a **scatterplot**. In a scatterplot, the ith data pair (x_i, y_i) is illustrated with a mark (such as a circle or cross) whose x-coordinate has the value x_i and whose y-coordinate has the value y_i.

In the following scatterplot, we have $n = 50$ datapoints for the 50 US states; x is the Republican share of the vote in the 2016 Trump–Clinton presidential election, and y is the Republican share of the vote 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**

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

and s_x and s_y are the standard deviations.

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

For the elections data, the correlation is

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

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

which, as we expected, is extremely high.

Summary

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

Problem Sheet 1

This is Problem Sheet 1, which covers material from Section 1 of the notes. You should work through all the questions on this problem sheet during week 1, in preparation for your tutorial in week 2. Questions C1 and C2 are assessed questions, and are due in by **2pm on Friday 8 September**. I recommend spending about 3 hours on this problem sheet in week 1, plus 1 extra hour in week 2 to neatly write up and submit your answers to the assessed questions.

A: Short questions

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

:::: {.myq} **A1.** Consider the following data sets of the age of elected politicians on a local council.

TABLE

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

B: Long questions

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

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

(a) Total numbers of Skittles colours in a large packet of six mini-packets:

("Skittles" are a small coloured fruit-flavoured candy, if it matters.)

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

(b) Shirt sizes for a university football team:

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

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

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

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

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

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

C: Assessed questions

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

The deadline for submitting your solutions is **2pm on Friday 8 September**. Submission will be via Gradescope; submission will open on Monday 3 September. You should submit your answers as a single PDF file. Most students choose to hand-write their work, then scan it to PDF using their phone; if you do this, you should use a proper scanning app (like Microsoft Lens or Adobe Scan) – please do not just submit photographs. We will discuss Gradescope submission further in the week 2 lectures. Your work will be marked by your tutor and returned on Monday 18 September, when solutions will also be made available.

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

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

C1.

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

Part II: Probability

Chapter 2

Probability spaces

2.1 What is probability?

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 (that is, 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 likely 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, given their own thoughts and 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.

  - This …

- **Mathematical probability:** We have a function that assigns to each event a number between 0 and 1, called its probability, and that function has to obey certain mathematical rules, called "axioms".

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

2.2 Sample space 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.

This will be easier to understand with some concrete examples.

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

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

Example 2.3. integers

Example 2.4. Suppose we want a computer to pick a random number between 0 and 1. We could take 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; 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.

For any sample space Ω, there are two special events that always exist. There's Ω itself, the event containing all of the sample outcomes, which represents "something happens". There's also the empty set $\emptyset$, which contains none of the sample outcomes, which represents "nothing happens". Common sense suggests that Ω should have probability 1, because *something* is bound to happen – this will later be one of our probability "axioms". Common sense also suggests that $\emptyset$ should have probability 0, because it can't be that *nothing* happens – this will not be one probability axioms, but we'll show that it follows logically from the axioms we do choose.

2.3 Basic set theory

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

set theory notation

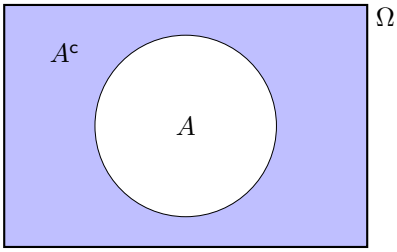
First, there are ways we can build new sets (or events) out of old.

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

- **NOT:** The **complement** of A, written A^{c} (and said "A complement" or "not A"), is the set of sample points not in A; that is

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

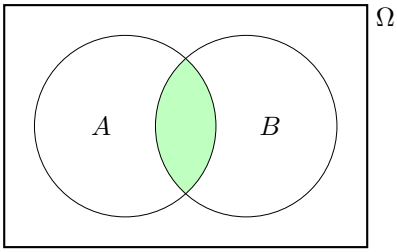
  This represents the event that A does not occur.



- **AND:** The **intersection** of A and B, written $A \cap B$ (and said "A intersect B" or "A and B") is the set of sample points in both A and B; that is,

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

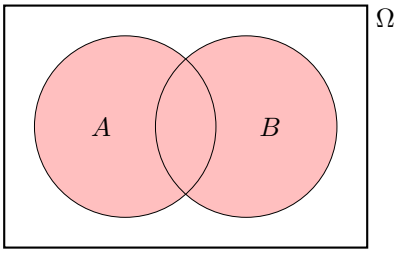
  This represents the event that both A and B occur.



- **OR:** The **union** of A and B, written $A \cup B$ (and said "A union B" or "A or B") is the set of sample points in A or in B; that is,

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

  This represents the event that A occurs or B occurs. (In mathematics, "or" includes "both", so a sample outcome in both A and B is in $A \cup B$ too.)



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

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

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.

De Morgan, etc.

2.4 Probability axioms

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

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

1. $\mathbb{P}(A) \geq 1$ for all events $A \subset \Omega$;
2. $\mathbb{P}(\Omega) = 1$;
3. if $A_1, A_2, ...$ is a finite or countably inifinite 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**.

HISTORICAL NOTE.

EXPLAIN AXIOMS.

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

But first, let's see some examples.

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

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

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

Let's check that the axioms hold:

1. Since $0 \leq p \leq 1$, all the probabilites are greater than or equal to 0.
2. It is indeed the case that $\mathbb{P}(\Omega) = \mathbb{P}(\{\mathrm{H}, \mathrm{T}\}) = 1$.

3. The only nontrivial disjoint union to check is $\{\mathrm{H}\} \cup \{\mathrm{T}\} = \{\mathrm{H}, \mathrm{T}\}$. But

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

as required.

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

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

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

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

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

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

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

2.5 Properties of probability

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

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

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

EXPLAIN?

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

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

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

Applying Axiom 3 to this disjoint union gives

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

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

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

and we are done. □

2.6 Addition rules for unions

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

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

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

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

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

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

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

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

You may have seen this result before. You've perhaps justified it by saying something like this: "We can add the two probabilities together, except now we've double-counted the overlap, so we have to take the probability of that away." That's OK as a way to remember the result – but this is a proper university mathematics course, so we want to prove it formally starting from just the axioms.

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

PICTURE

Proof. First note that we have

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

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

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

The left-hand side looks good, and the first term on the right-hand side looks good. To deal with the second term on the right-hand side, we need to write it down as part of a disjoint union again. Can we find another one? Yes! We have

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

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

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

Rearranging this gives

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

PICTURE

Finally, substituting (2.2) into (2.1) gives

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

as required. □

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

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

We have $\mathbb{P}(\text{spade}) = \frac{13}{52} = \frac{1}{4}$ and $\mathbb{P}(\text{Ace}) = \frac{4}{52} = \frac{1}{13}$. So we have

$$\mathbb{P}(\text{spade or Ace}) = \tfrac{1}{4} + \tfrac{1}{13} - \mathbb{P}(\text{spade and Ace}).$$

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

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

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

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

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

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

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

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

Chapter 3

Classical probability

3.1 Probability with equally likely outcomes

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

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

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

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

So to work out a classical probability, we need to be able to count how many outcomes are in A and count how many outcomes are in Ω. In this section, we'll see a number of situations where we can do this.

PROOF SOMEWHERE

EASY EXAMPLE

3.2 Multiplication principle

In classical probability, to find the probability of an event A, we need to count the number of outcomes in A and the total number of possible outcomes in

Ω. This can be easy when we're just looking at one choice – like the XXXXX YYYYY above. 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**. This says that if you have n choices followed by m choices, than all together you have $n \times m$ total choices. You can see this by imagining the choices in a $n \times m$ grid, with the n columns representing the first choice and m rows representing the second choice. For example, suppose you go to a burger restaurant where there are 3 choices of burger (beefburger, chicken burger, veggie burger) and 2 choices of sides (fries, salad), then altogether there are $3 \times 2 = 6$ choices of meal.

TABLE

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

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

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

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

So the probability all five coins show the same face is

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

3.3 Sampling with and without replacement

3.4 Sampling without replacement and without labelling

3.5 Birthday problem

Problem Sheet 2

Chapter 4

Independence and conditional probability

Chapter 5

Discrete random variables

Problem Sheet 3

Chapter 6

Discrete distributions

Chapter 7

Multiple random variables

Problem Sheet 4

Chapter 8

Continuous random variables

Chapter 9

Normal distribution

Part III: Bayesian statistics

Chapter 10

Introduction to Bayesian statistics

Other stuff

Problem Sheet 5

Chapter 11

Summary

R Worksheets

R worksheets

Each week there will be an R worksheet to work through in your own time. We recommend spending about one hour on each worksheet, plus one extra hour for worksheets with assessed questions, for checking through and submitting your solutions.

Week	Worksheet	Deadline for assessed work
1	R basics	—
2	Working with vectors	—
3	Importing data into R	Friday 15 October
4	Plots I: Making plots	—
5	Plots II: Making plots nicer	Friday 29 October
6	RMarkdown (optional)	—
7	Discrete random variables	Friday 12 November
8	Discrete distributions	—
9	Normal distribution	Friday 26 November
10	Law of large numbers	—
11	Summary	Friday 10 December

About R and RStudio

- **R** is a *programming language* that is particularly good at working with probability and statistics. R is very widely used in universities and increasingly widely used in industry. Learning to use R is a mandatory part of this module, and exercises requiring use of R make up at least 15% of your module mark. Many other statistics-related course at the University also use R.
- **RStudio** is a *program* that gives a convenient way to work with the language R. RStudio is the most common way to use the language R, and learning to use RStudio is strongly recommended.

R and RStudio are free/open-source software.

How to access R and RStudio

There are a number of ways you can access R and RStudio:

- All **University computers** have R and RStudio already installed. Here is a directory of the University's computer clusters.
- You can **install** R and RStudio on your own computer – see the instructions below.
- If you want to use R/RStudio on a non-University device for which you don't have admin/installation rights (Chromebook, iPad, friend's laptop, etc), you could try:
  - You can use the University's copies of R/RStudio virtually through the Windows Virtual Desktop or AppsAnywhere client.
  - The RStudio Cloud is a cloud-hosted "Google Docs for R" that you can use through your web browser – you can get 15 hours per month for free (or pay for more).

Installing R and RStudio

Students who have their own computer usually find it most convenient to install R and RStudio on that computer. To do this, it's important that you install R (the programming language) first, and only install RStudio (the program to use R) once R has already been installed.

1. *First*, install **R**. Go to the Comprehensive R Archive Network and follow the instructions:
   - Windows: Click "Download R for Windows", then "Install R for the first time". The main link at the top should be to download the most recent version of R.
   - Mac: Click Download R for macOS, and then download the relevant PKG file. (For pre-November 2020 Intel-based Macbooks, you must use the "Intel 64-bit build"; for post-November 2020 M1-based "Apple silicon" Macbooks, the "Apple silicon arm64 build" may be faster.)
2. *After* R is installed, *then* install **RStudio**. Go to the Download page at RStudio.com and follow the instructions. You want "RStudio Desktop", and you want the free version.

If you have difficulty installing R, come along to the first computational drop-in session in Week ?? and bring your computer with you (if it's sufficiently portable), and we'll do our best to help.

11.1 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 get frustrated and annoyed when the computer seems to refuse to do what you want it to. 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 XXXXXXXx