MATH1710 Probability and Statistics I

Matthew Aldridge

University of Leeds, 2023–24

Contents

Schedule

Week 3 (16–20 October):

- **Lecture 5:** Classical probability I (Monday 16 October)
- **Lecture 6:** Classical probability II (Wednesday 18 October)
- **R Practical**
- **Problem Sheet 2:** Work through in preparation for your tutorial in Week 4. Deadline for assessed questions: Monday 30 October.
- **R Worksheet 2:** Deadline for assessed exercises: Monday 23 October
- **Office hours:** Friday 20 October, 11–12 and 1–2, Maths Boardroom

Week 2 (9–13 October):

- **Lecture 3:** Sample spaces and events (Monday 9 October)
- **Lecture 4:** Probability (Wednesday 11 October)
- **Tutorial** on Problem Sheet 1
- **R Practical**
- **Problem Sheet 1:** Work through in preparation for your tutorial. Deadline for assessed questions: Monday 16 October.
- **R Worksheet 1**
- **Office hours:** Friday 13 October, 11–12 and 1–2, Maths Boardroom

Week 1 (2–6 October):

- **Lecture 1:** Summary statistics (Monday 2 October)
- **Lecture 2:** Data visualisation (Wednesday 4 October)
- **Problem Sheet 1:** Work through in preparation for your tutorial in Week 2. Deadline for assessed questions: Monday 16 October.
- **Office hours:** Friday 6 October, 11–12 and 1–2, Maths Boardroom

About MATH1710

Organisation of MATH1710

This module is **MATH1710 Probability and Statistics I**. (A small number of second-year scientists are taking this module as the first half of **MATH2700 Probability and Statistics for Scientists**.)

This module lasts for 11 weeks from 2 October to 15 December 2023. The exam will take place between 15 and 26 January 2024.

The module leader, the lecturer, and the main author of these notes is Dr Matthew Aldridge (you can call me "Matt", "Matthew", 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 each lecture will be delivered twice:

- **Mondays** at 1200 or at 1400, in Chemistry West LT F
- **Wednesdays** at 1500 in Chemistry West LT F or at 1600 in Roger Stevens LT 20

Check your timetable to see which lecture you are assigned to each day (or click this link for a map of how to get to Chemistry West LT F.

I recommend taking your own notes during the lecture. I will put brief summary notes from the lectures on this website, 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 work on 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 recommend completing and submitting the work in the previous week.

Problem Sheet	Lectures covered	Deadline for assessed work
1	1 and 2	Monday 16 October (Week 3)
2	3–6	Monday 30 October (Week 5)
3	7–10	Monday 13 November (Week 7)
4	11–14	Monday 27 November (Week 9)
5	15–18	Monday 11 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 online through the Gradescope platform. Most students choose to hand-write their solutions on paper and then scan and submit on their phone using the Gradescope app. Further Gradescope details to follow nearer the first deadline.

Tutorials

Tutorials are small groups of about a dozen students. You have been assigned to one of 34 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 tutorials are an interactive session, where the main goal will be to go over your answers to the non-assessed questions on the problems sheets, which you will have worked on in advance of the tutorial. 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 without 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. 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 and Practicals

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, including 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, starting from Week 2. Even-numbered worksheets 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
2	1: R basics	—
3	2: Vectors	Monday 23 October (Week 4)
4	3: Data in R	—
5	4: Plots I – Making plots	Monday 6 November (Week 6)
6	5: Plots II – Making plots better	—
7	6: Discrete distributions	Monday 20 November (Week 8)
8	7: Discrete random variables	—
9	8: Normal distribution	Monday 4 December (Week 10)
10	9: Law of large numbers	—

Week	Worksheet	Deadline for assessed work
11	10: Recap	Thursday 14 December (Week 11)

R Practical sessions: You will be introduced you to R in your first Practical session, in Week 2. You will first see how to use R on University computers (these sessions will take place in computer "clusters"). There will then be an opportunity to install R on your own device – if you have a laptop on which you want to install R, bring it along to the practical session. A second practical, in Week 3, will allow you to get help on the R Worksheet 2, which is the first worksheet with assessed questions.

There are 11 R practical session groups – check your timetable for Weeks 2 and 3 to see when and where your group meets.

Attendance at the first R practical session (Week 2) is compulsory.

"Office hours" drop-in sessions

If you there is something in the module you wish to discuss 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 me; 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 "office hours" drop-in sessions per week:

- Fridays 1100–1200 and 1300–1400 in the Mathematics Boardroom (map).

I may change arrangements as term continues – if attendance levels are low, I will move office hours to be actual office.

If neither time is possible, you may email me to arrange an alternative time to talk to me.

Attendance at "office hours" sessions is optional.

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:** 16 hours total at the end of the module.
- **Exam:** 2 hours.

That makes about 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 440 students registered for 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 15 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 ask at your R practical 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 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. (Extensions of up to 4 days are available for Problem Sheets. Only exemptions 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 many 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", but uses probabilistic techniques developed here.

Books

You can do well on this module by 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 pre-reading in advance, 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 a lot in 2022 and a little bit in 2023. 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 module, 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:

```
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. 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 6, 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. (If Sigma notation is new to you, I recommend this PDF from MathCentre, or Section 2.4 of Clarke and Cooke, *A Basic Course in Statistics*.)

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 $\text{IQR} = q(\frac{3}{4}) - q(\frac{1}{4})$.

The **sample variance** is

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

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

Example 1.2. We continue with the Skittles data.

The number of distinct observation is 5. (These are 59, 60, 61, 62, and 63.)

The interquartile range is $x_{10} - x_4 = 62 - 59 = 3$.

You will calculate the sample variance on Problem Sheet 1.

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.

Recommended reading:

- Wikipedia: Exploratory data analysis, Mode (statistics), Median, Arithmetic mean, Quantile, Interquartile range.
- Clarke and Cooke, *A Basic Course in Statistics*, Sections 2.1–2.4, 2.7, 4.1–4.4, 4.6, 4.7.

On Problem Sheet 1, you should now be able to complete Questions A1, B1, B2, B4, C1.

Chapter 2

Data visualisations

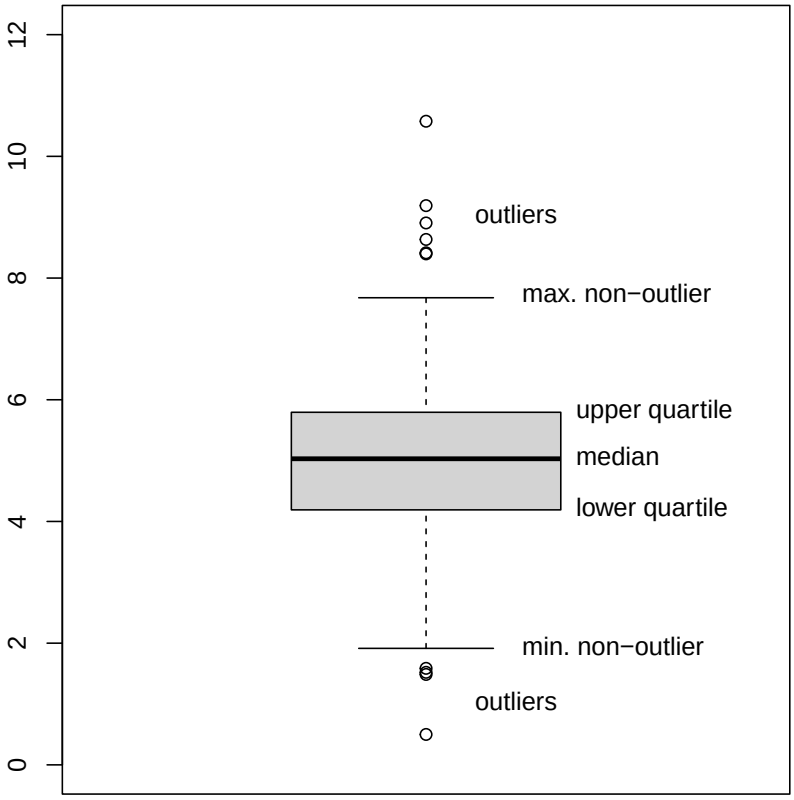
Data visualisations – drawings or graphs based on data – can help us to understand the "shape" of a dataset as part of exploratory data anlaysis. In this lecture, we'll look at three types of data visualisation.

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.

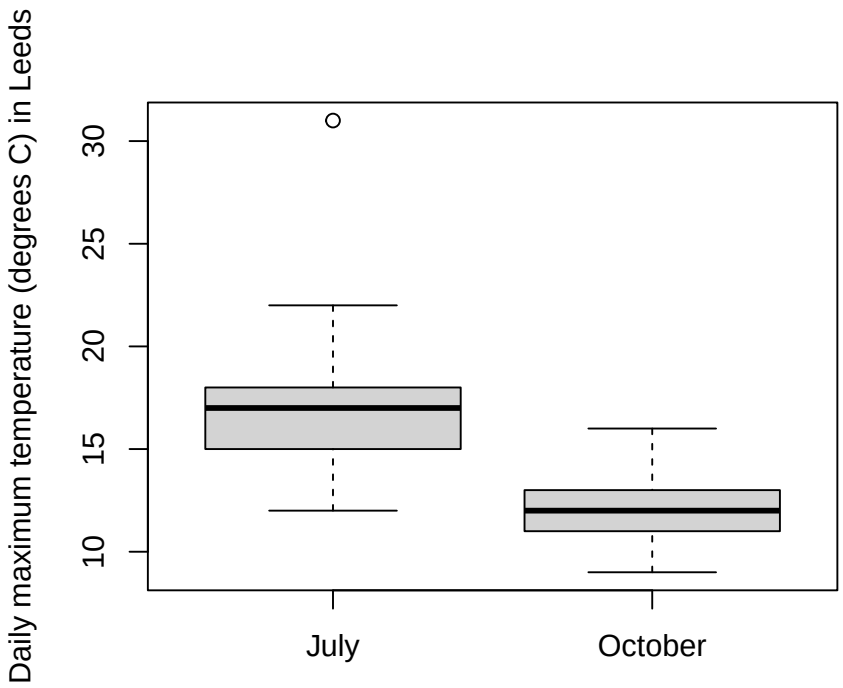
A boxplot is drawn as follows:

- 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 $q(\frac{1}{4}) - 1.5 \times \text{IQR}$ 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 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	

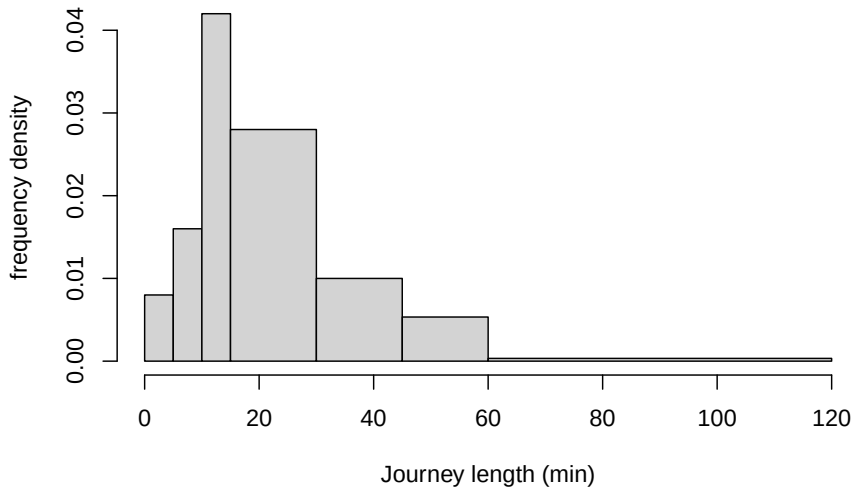
In the first row, for example, the relative frequency is 0.04 and the size of the bin is 5 minutes, so the frequency density is $0.04/5 = 0.008$. We now see that the modal bin – the bin with the highest frequency *density* – is in fact the "10–15 minutes" bin. This bin has somewhat fewer datapoints that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

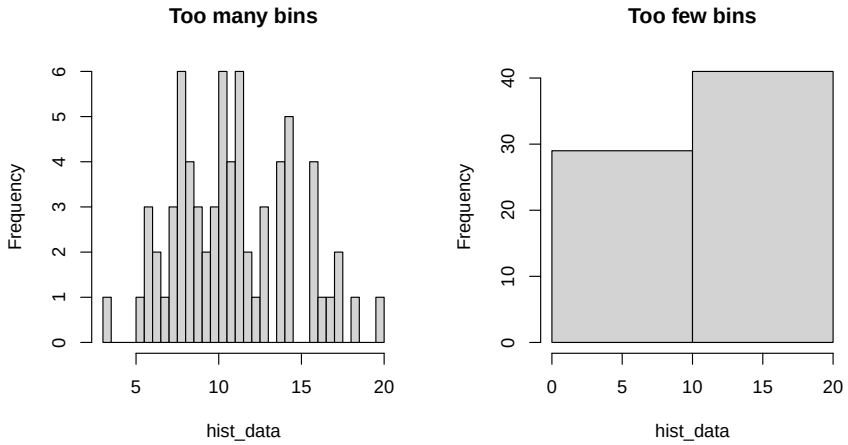
Here is a histogram for our journey-time data:

```
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

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

2.3 Scatterplots

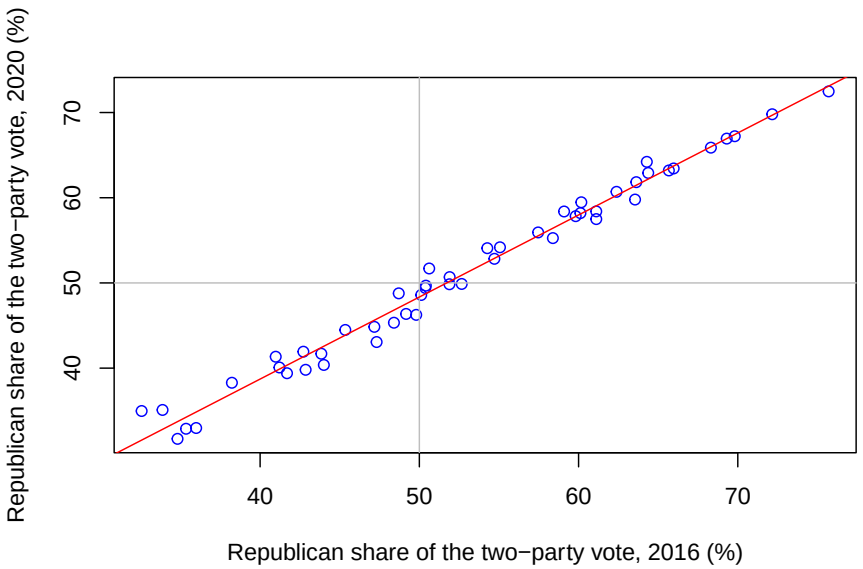
Often, more than one piece of data is collected from each subject, and we wish to compare that data, to see if there is a relationship between the variables.

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, $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$. 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 $\mathbf{x}$ and $\mathbf{y}$, 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**

$$\begin{aligned} s_{xy} &= \frac{1}{n-1}((x_1 - \bar{x})(y_1 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})) \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), \end{aligned}$$

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.

Recommended reading:

- Wikipedia: Box plot, Histogram, Grouped data, Scatter plot, Pearson correlation coefficient.
- Clarke and Cooke, *A Basic Course in Statistics*, Sections 1.2, 2.5, 4.5, 4.6, 21.2, 21.3.

On Problem Sheet 1, you should now be able to complete all questions.

Problem Sheet 1

Solutions are now available to all non-assessed questions.

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 advance of your tutorial in Week 2. Questions C1 and C2 are assessed questions, and are due in by **2pm on Monday 16 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 after Friday 13 October. 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 definitional formula.

(c) Calculate the sample variance using the computational 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		
30–40	2		
40–45	4		
45–50	5		
50–60	6		
60–80	2		
Total	20	1	—

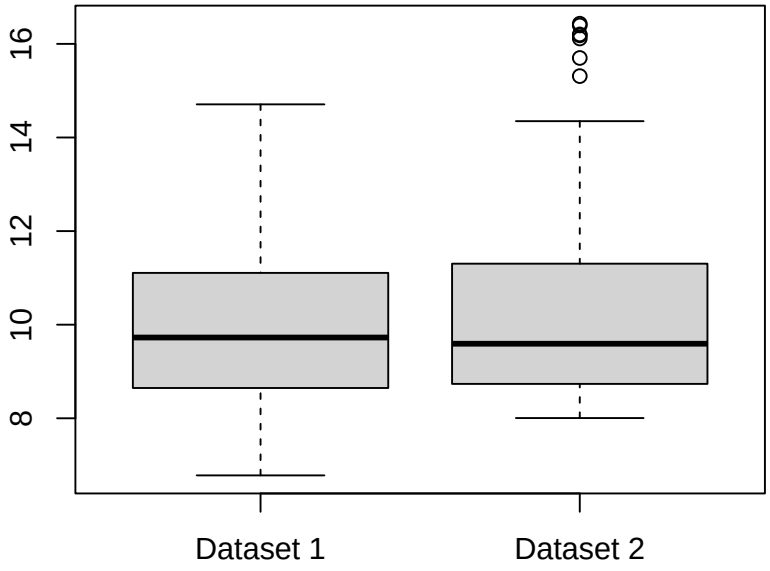
(a) Complete the table by filling in the relative frequency and frequency densities.

(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 write good, clear solutions – are likely to be the main topic for discussion in your tutorial. Solutions will be available after Friday 13 October.

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) Shirt sizes for the $n = 16$ members of a university football squad:

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

(b) Six packets of Skittles are opened together, a total of $n = 361$ sweets. The colours of these sweets is recorded as follows:

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

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 16 October** at the beginning of Week 3. Submission will be via Gradescope, which you can access via Minerva or on the Gradescope mobile app. You should submit your answers as a single PDF file. Most students choose to hand-write their work on paper, then scan-and-submit it to using the Gradescope mobile app. Your work will be marked by your tutor and returned on Monday 23 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 explanatory writing.

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

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

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

(a) Calculate the median for this data.

(b) Calculate the mean for this data.

(c) Calculate the sample variance for this data.

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

C2. **(a)** 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.

(b) Prove the following computational formula for the sample covariance:

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

Solutions to short questions

A1. (a) 60.7, (b) 2.40, (c) 2.40, (d) —.
A2. (a) —, (b) 45–50, (c) 45–50, (d) 48.6 *(corrected)*.
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 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 Labour win the 2024 general election?"). 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 outcomes 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 outcomes not in A; that is

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

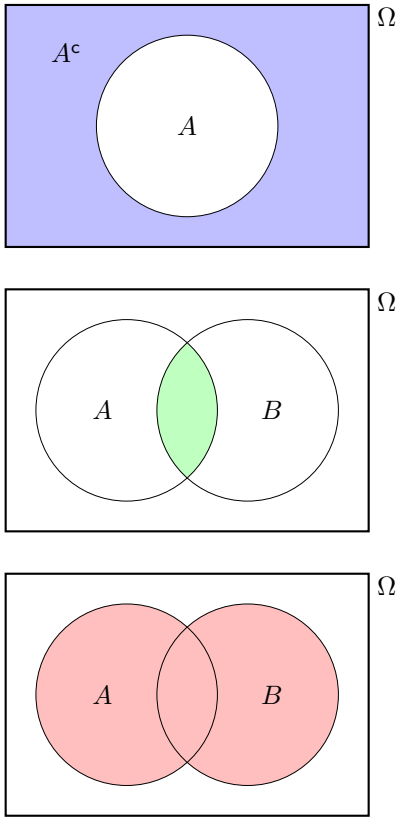
  This represents the event that A does not occur.
- **AND:** The **intersection** of A and B, written $A \cap B$ (and said "A intersect B" or "A and B") is the set of sample outcomess in both A and B; that is,

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

  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 outcomess 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

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

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:

$$(A^c)^c = A.$$

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:

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

  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.)

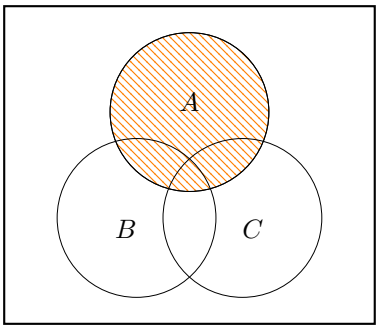
For this module, these mostly count as "common sense" – but if you ever do need to prove one of these statements (or a similar one), one way is to use a Venn diagram.

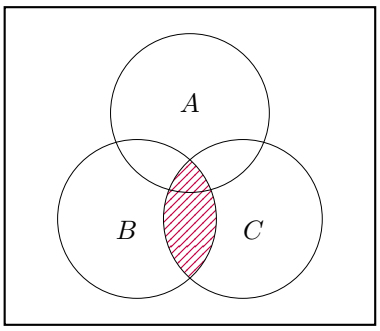
Let's prove the second distributive law,

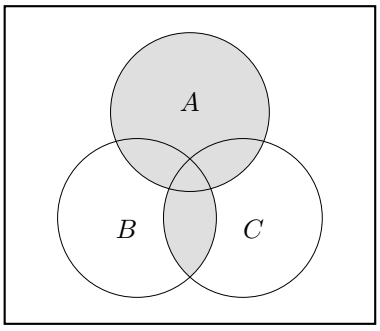
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C),$$

with a Venn diagram as an example.

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

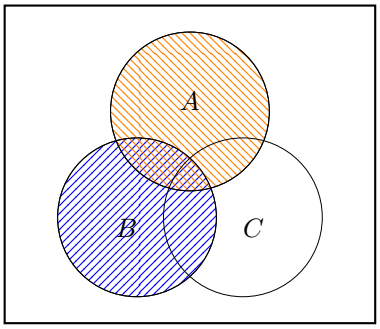


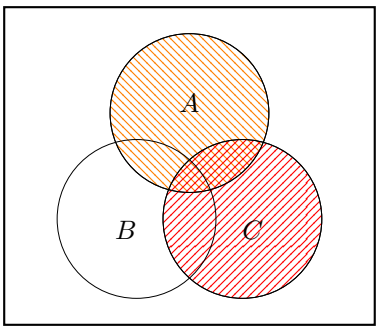


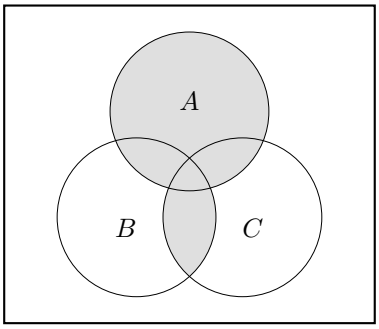


The left-hand figure is A, the middle figure is $B \cap C$, and the right-hand figure is union of these, $A \cup (B \cap C)$.

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" $A \cup B$.

Recommended reading:

- Stirzaker, *Elementary Probability*, Sections 1.1 and 1.2 (plus optionally Chapter 0).
- Grimmett and Welsh, *Probability*, Sections 1.1 and 1.2.

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 is about *disjoint* events – recall that these are events where no two can happen at the same time, because the intersection of any pair of them is empty. [***Correction:*** *In the 3pm lecture, I wrongly said that their* union *is empty.*] 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 only 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

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

Let's check that the axioms hold:

1. Since $0 \leq p \leq 1$, all the 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.

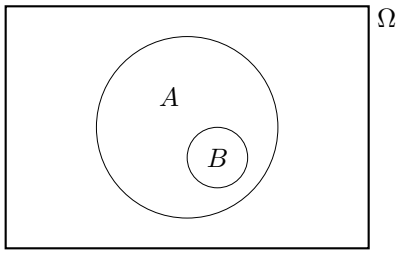
Theorem 4.1. *Let Ω be a sample space with a probability measure $\mathbb{P}$. Then we have the following:*

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

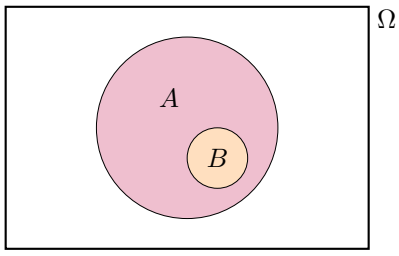
Proof. The key with most of these "prove from the axioms" problems is to think of a way to write the relevant events as part of a *disjoint* union, then use Axiom 3. Statements 1 and 2 are exercises for you on Problem Sheet 2. We'll start with the third statement.

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

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



Applying Axiom 3 to this disjoint union gives

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

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 $A_1 = A$, $A_2 = B$. □

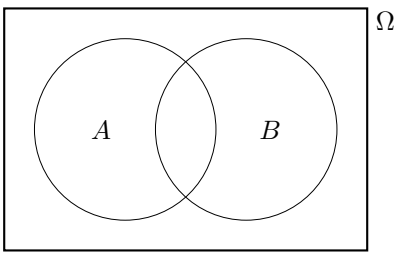
But what about if A and B are not disjoint? Then we have the following.

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

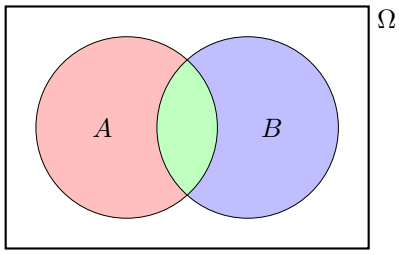
$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

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.

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" $A \cap B^c$, "B but not A" $B \cap A^c$, and "both" $A \cap B$.



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

$$A = (A \cap B^c) \cup (A \cap B) \tag{4.1}$$
$$B = (B \cap A^c) \cup (A \cap B) \tag{4.2}$$
$$A \cup B = (A \cap B^c) \cup (B \cap A^c) \cup (A \cap B), \tag{4.3}$$

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

$$\mathbb{P}(A) = \mathbb{P}(A \cap B^c) + \mathbb{P}(A \cap B) \tag{4.4}$$
$$\mathbb{P}(B) = \mathbb{P}(B \cap A^c) + \mathbb{P}(A \cap B) \tag{4.5}$$
$$\mathbb{P}(A \cup B) = \mathbb{P}(A \cap B^c) + \mathbb{P}(B \cap A^c) + \mathbb{P}(A \cap B). \tag{4.6}$$

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

$$\mathbb{P}(A \cap B^c) = \mathbb{P}(A) - \mathbb{P}(A \cap B) \tag{4.7}$$
$$\mathbb{P}(B \cap A^c) = \mathbb{P}(B) - \mathbb{P}(A \cap B). \tag{4.8}$$

Substituting these into (4.6) gives

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) - \mathbb{P}(A \cap B) + \mathbb{P}(B) - \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B) \tag{4.9}$$
$$= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B), \tag{4.10}$$

as required. □

Example 4.3. *Consider picking a card from a standard 52-card 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 $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

Recommended reading:

- Stirzaker, *Elementary Probability*, Sections 1.3 and 1.4.
- Grimmett and Welsh, *Probability*, Sections 1.3 and 1.4.

On Problem Sheet 2, you should now be able to complete Questions A1, A2, B1, B2 and perhaps C1.

Chapter 5

Classical probability I

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.

Recommended reading:

- Stirzaker, *Elementary Probability*, Sections 3.1 and 3.2.

Chapter 6

Classical probability II

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 $\binom{n}{k} = n^{\underline{k}}/k!$.

Recommended reading:

- Stirzaker, *Elementary Probability*, Sections 3.2 and 3.3.

On Problem Sheet 2, you should now be able to complete all questions.

Problem Sheet 2

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 30 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) $\binom{10}{4}$

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

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 $\{1, 2, \ldots, 2023\}$.

(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 $\{0, 1, \ldots, 9\}$, each with equal probability. A statistician chooses n random digits (with replacement).

(a) For $k = 0, 1, \ldots, 9$, 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 30 October** at the beginning of Week 5. You should submit a PDF to Gradescope. Your work will be marked by your tutor and returned on Monday 6 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 explanatory writing.

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) $\{0, 1, \ldots, 4\}$ (ii) $\{\text{HHHH}, \text{HHHT}, \text{HHTH}, \ldots, \text{TTTT}\}$ (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) $\frac{4}{25} = 0.16$ (b) $\frac{2}{15} = 0.133$

Other stuff

R Worksheets

Each week (starting in Week 2) 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 even-numbered worksheets with assessed questions, for checking and submitting your solutions.

Week	Worksheet	Solutions	Deadline for assessed work
2	**1: R basics**	Solutions	—
3	**2: Vectors**		Monday 23 October

R Practical 1

R Worksheet 1 is here

About the Practical

The first computer practical sessions are this week to introduce you to the R programming language. You should make sure:

- You know where and when your practical session is – check your timetable!
- You know your username, password, and the Duo authentication system to log on to university computers.
- To bring your laptop along if (optionally) you want to install R and RStudio on it.

Below are some directions of how to find the various computer clusters.

What are R and RStudio?

- **R** is a *programming language* that is particularly useful for working with probability and statistics. The R language 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 *computer program* (or *app*) that gives a convenient way to work with the language R. The RStudio program is made by the company Posit. THe program 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.

There are a number ways you can use R and RStudio:

1. *On University computers.* You will learn how to use R and RStudio on University computers in your first practical session, in Week 2.

2. *On your own computer.* R and RStudio can easily be installed on Windows and Mac laptops. Bring your laptop along to the first practical session to learn how to install R and RStudio.
3. *Using the Posit Cloud.* The Posit Cloud is a way to use R and RStudio online – sort of like a "Google Docs for R". You can use it free for 25 hours a month, which should be plenty for this module, or pay for more. I recommend the Posit Cloud for using R/RStudio with Chromebooks, tablet computers, or when borrowing someone else's device.

Accessing R and RStudio on University computers

R and RStudio can be used on University computers via the AppsAnywhere portal. AppsAnywhere is the University of Leeds system for loading "unusual" programs (common programs like Microsoft Office and web browsers are preloaded).

There are three steps to using R and RStudio on University computers:

1. Open the AppsAnywhere portal.
2. Load the language R onto your computer.
3. Open the program RStudio.

First, open the AppsAnywhere portal by double-clicking on the desktop icon. This will open a web browser, and invite you to "Open AppsAnywhere Launcher" – you should accept and open. AppsAnywhere has loaded properly when the blue "Validation in progress…" box turns into a green "Validation Successful" box.

Second, launch R from AppsAnywhere. R is called "Cran R 4.2.0 x64" on AppsAnywhere, so searching for "Cran" is an easy way to find it. Click "Launch".

This will do two things. First, it will silently load the language R in the background. Second, it will open a program called "RGui". RGui is basically like an older and less good version of RStudio; we do not recommend using the RGui program, so you can close it. (The R language will remain loaded.)

Third, launch RStudio from AppsAnywhere. The most recent version on AppsAnywhere is "RStudio 2023 (03.0.386)". Click "Launch". After a few second, RStudio will launch. (If invited to choose a version of the language R, pick "64-bit". If invited to update R or RStudio, decline.)

You need to repeat all three steps each time you log onto a University computer.

Installing R and RStudio

When you install R and RStudio, it's important that you install the language R first, and only install the program RStudio after the language 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. (Most modern Macbooks are based on Apple's M1 or M2 processors, so you can choose "Apple silicon arm64 build". Some older Macbooks, mostly 2020 or earlier, have Intel processors; for these you should use the "Intel 64-bit build".)

2. *After* R is installed, *then* install RStudio. Go to the "Download RStudio" page at posit.co and follow the instructions. You want "RStudio Desktop" and you want the free version, if given a choice.

Now, whenever you want to use R and RStudio, simply open program RStudio. (The language R will automatically be loaded on your computer.)

For Chromebooks, we recommend using the Posit Cloud, as mentioned above. However, if you have an Intel-based Chromebook and are feeling brave, then we have had success installing R and RStudio using these instructions, which are long and complicated.

Where are the computer clusters?

EC Stoner Cluster 6.68

EC Stoner Cluster 6.68 [map] is in the **EC Stoner building**. The easiest way to enter the 6th floor of EC Stoner is via the sliding doors ("South Entrance 4") opposite the multi-storey car park (red arrow).

Alternatively, from the "red route" along the 10th floor corridor of EC Stoner, go all the way along to the School of Food Science and Nutrition, by staircase 4, and take the lift (*not* the stairs) to the 6th floor (blue dotted line).

EC Stoner Cluster 6.68 is the big room to the east (right from the doors; left from the lift), through the smaller ante-room.

Fourman O & P Clusters

The **Fourman O & P Clusters** [map] are in the **Worsley building**, at the south of campus. The easiest route is from Chancellor's Court, next to the Roger Stevens building: follow the North–South Access Route alongside and then through the Garstang building, and enter the Worsley building at the 7th floor "airport lounge" area (red arrow). Follow the signs to "Central Teaching Space": turn left to get to the East staircase, and go up one floor the the 8th floor.

Alternatively, enter the Worsley building on Clarendon Way at the 4th floor (blue arrow). The East lifts are to your left: take the lift to the 8th floor.

On the 8th floor of the Worsley building, the Fourman Clusters are two connected rooms right next the East staircase/lifts – Fourman O is on the right; Fourman P is on the left.

Irene Manton North & South Clusters

The **Irene Manton Clusters** [map] are in the **Irene Manton building**, which is behind the Roger Stevens building. You can get there out the back door of Roger Stevens, by LT 01, or via the gardens to the right of Roger Stevens.

Walking down the walkway alongside the Irene Manton building, there's a modest, unassuming door on your left. Enter it. Irene Manton North is the room to your left; Irene Manton South is to your right.

Psychology Cluster 1.43

Psychology Cluster 1.43 [map] is in the **Psychology building**, which is in between the two main east–west thoroughfares through campus – University Road and Beach Grove Terrace/Precinct/Lifton Place. One entrance to Psychology is on Lifton Place, opposite Cromer Street, up some steps; the other is on University Road, opposite the School of Fine Art, through a little garden.

Once inside Psychology, there are signs to the "CBL clusters". Follow a long winding corridor to the east (right from Lifton Place entrance; left from University Road entrance), past a giant glittery brain sculpture, all the way to a spiral staircase. Go up one floor. Psychology Cluster 1.43 is the room on the right.

Richard Hughes Cluster

The **Richard Hughes Cluster** [map] is on the side of the **Clothworkers' Link building**, on the north side of University Road by the "Clothworkers' Link" – the sky-bridge that goes over University Road.

Just to the west (Hyde Park side) of the Link bridge, go up the alleyway between the Clothworkers' Link building and 28 University Road. Enter a small door on the right; the Richard Hughes Cluster is at the top of a short flight of stairs to your left.

Solutions

This page has the solutions to all the non-assessed questions on Problem Sheet 1. Solutions are added after all 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

A: Short questions

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:

$$\bar{x} = \frac{1}{13}(59 + 59 + \cdots + 63) = \frac{789}{13} = 60.6923\cdots \approx 60.7.$$

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

Solution.

$$\begin{aligned} s_x^2 &= \frac{1}{13-1}\left((59-60.7)^2+(59-60.7)^2+\cdots+(63-60.7)^2\right) \\ &= \frac{1}{12}(2.86+2.86+\cdots+5.33) \\ &= \frac{1}{12}\times 28.77 \\ &= 2.40 \end{aligned}$$

(c) 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 $\bar{x}$.

(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 different answers 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	2		
40–45	4		
45–50	5		
50–60	6		
60–80	2		

Age (years)	Frequency	Relative frequency	Frequency density
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	0.0042
30–40	2	0.1	0.01
40–45	4	0.2	0.04
45–50	5	0.25	0.05
50–60	6	0.3	0.03
60–80	2	0.1	0.005
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.

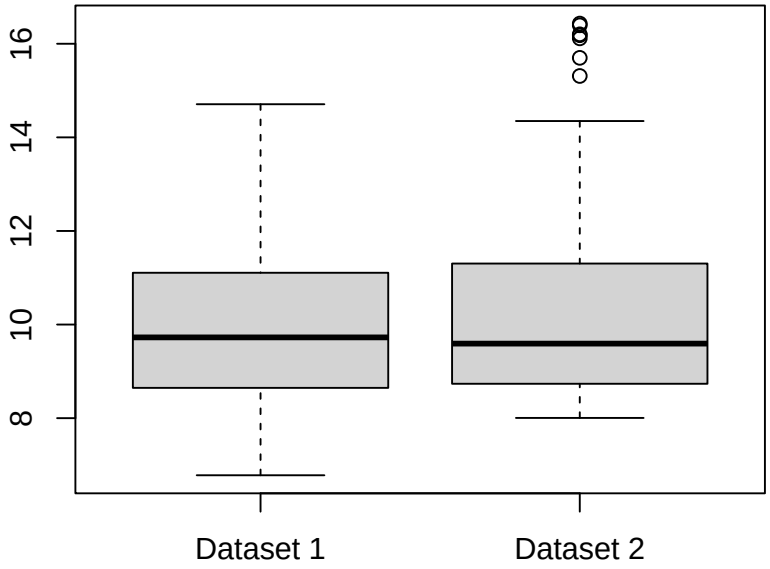
Group feedback: Remember that the modal bin is the one with the largest frequency *density*, not necessarily the bin with the highest frequency.

(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

$$\bar{x} = \frac{1}{20}(1 \times 24 + 2 \times 35 + \cdots + 2 \times 65) = \frac{971.9}{20} = 48.6.$$

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.

B: Long questions

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) Shirt sizes for the $n = 16$ members of 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.

(b) Six packets of Skittles are opened together, a total of $n=361$ sweets. The colours of these sweets is recorded as follows:

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.

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: the 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) An outlier will typically be the only data point with its value, or certainly rare. Therefore, 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 any 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. This is not typically a relevant increase, so the number of distinct outcomes 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 (owner of Tesla and Twitter, net worth roughly $200 billion) joined my tutorial group? The mean and sample variance would change an enormous amount, while the median and interquartile range would barely change at all in comparison.

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 $\mathbf{a} = (a_1, a_2, \ldots a_n)$ 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

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

(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,

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

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

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2,$$

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 that 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 $s_{xy}^2 \leq s_x^2 s_y^2$, as required.

Group feedback: Keep trying different choices for (a_i) and (b_i); maybe your first attempt won't work, but it pays to be persistent!

A fancier choice is $a_i = (x_i - \bar{x})/\sqrt{n-1}$ and $b_i = (y_i - \bar{y})/\sqrt{n-1}$, to get the exact result without needing a second cancellation step, but I would find that harder to spot.

(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; numbers smaller than -1 become bigger than +1 when squared; numbers between -1 and +1 get closer to 0.) 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 extremely careful if you do this – make sure you are properly accounting for the positive and negative square roots on both sides (if necessary), and where that does or doesn't require reversing the direction of 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 maintain privacy 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 publicly 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.