

# MATH1710 Probability and Statistics I

Matthew Aldridge

University of Leeds, 2021–22



# Contents

<b>Schedule</b>	<b>5</b>
<b>About MATH1710</b>	<b>7</b>
Organisation of MATH1710 . . . . .	7
Content of MATH1710 . . . . .	12
About these notes . . . . .	13
 <b>Part I: Exploratory data analysis</b>	 <b>17</b>
<b>1 Summary statistics</b>	<b>17</b>
1.1 What is EDA? . . . . .	17
1.2 What is R? . . . . .	18
1.3 Statistics of centrality . . . . .	19
1.4 Statistics of spread . . . . .	20
Summary . . . . .	21
 <b>2 Data visualisations</b>	 <b>23</b>
2.1 Boxplots . . . . .	23
2.2 Histograms . . . . .	25
2.3 Scatterplots . . . . .	28
Summary . . . . .	30
 <b>Problem Sheet 1</b>	 <b>31</b>
A: Short questions . . . . .	31
B: Long questions . . . . .	33
C: Assessed questions . . . . .	34
Solutions to short questions . . . . .	35

<b>R Worksheets</b>	<b>37</b>
R worksheets . . . . .	37
About R and RStudio . . . . .	37
How to access R and RStudio . . . . .	38
R troubleshooting drop-in sessions . . . . .	39

# Schedule

**Week 1** (3–7 October):

- **Lecture 1:** Summary statistics (Monday 3 October)
- **Lecture 2:** Data visualisation (Wednesday 5 October)
- **Problem Sheet 1:** discussed in tutorials in Week 2; deadline for assessed questions: Monday 17 October
- **R Worksheet 1:** R basics



# About MATH1710

## Organisation of MATH1710

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

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

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

## Lectures

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

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

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

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

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

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

*Attendance at lectures is compulsory.*

## Problem sheets

There will be 5 problem sheets. Each problem sheet has a number of short and long questions for you to cover in your own time to help you learn the material, and two assessed questions, which you should submit for marking. The assessed questions on each problem sheet make up 3% of your mark on this module, for a total of 15%. Deadlines are 2pm on Mondays, although I'd personally recommend completing and submitting the work in the previous week.

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

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

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

## Tutorials

Tutorials are small groups of about a dozen students. You have been assigned to one of 38 tutorial groups, each with a member of staff as the tutor. Your tutorial group will meet five times, in Weeks 2, 4, 6, 8, and 10; you should check your timetable to see when and where your tutorial group meets.

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

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



- Work through the problem sheet before the tutorial, spending plenty of time on it, and making multiple efforts at questions you get stuck on. I recommend spending *at least 4 hours per problem sheet*. This is a long time, but you shouldn't expect to be able to answer the hardest questions on a problem sheet with making multiple attempts. You don't have to wait until all lectures in a section are complete until starting to work on some of the questions – this is particularly important for students with Monday tutorials. Collaboration is encouraged when working through the non-assessed problems, but I recommend writing up your work on your own; answers to assessed questions must be solely your own work.
- Take advantage of the small group setting of the tutorial to ask for help or clarification on questions you weren't able to complete.
- After the tutorial, attempt again the questions you were previously stuck on.
- If you're still unable to complete a question after this second round of attempts, *then* consult the solutions.

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

*Attendance at tutorials is compulsory.*

## R worksheets

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

You will learn R by working through one R worksheet each week in your own time. Worksheets 3, 5, 7, 9 and 11 will also contain a few questions for assessment, which will be due by 2pm Monday the following week (except the last one). Each of these is worth 3% of your mark for a total of 15%. You will submit your answers through a Microsoft Form (details to follow later). I recommend spending one hour per week on the week's R worksheet, plus one extra hour if there are assessed questions that week.

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

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

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

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

*Attendance at R troubleshooting drop-in sessions is optional.*

### Optional “office hours” drop-in sessions

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

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

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

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

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

### Time management

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

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

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

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

## Exam

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

## Who should I ask about...?

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

- *I don't understand something in the notes or on a problem sheet:* Come to office hours, or ask your tutor in your next tutorial.
- *I'm having difficulties with R:* In Weeks 2 or 3, you should attend an R trouble-shooting drop-in session; at other times, come to office hours.
- *I have an admin question about arrangements for the module:* Come to office hours or talk to me before/after lectures.
- *I have an admin question about arrangements for my tutorial:* Contact your tutor.
- *I have an admin question about general arrangements for my programme as a whole:* Contact the Student Information Service or speak to your personal academic tutor.
- *I have a question about the marking of my assessed work on the problem sheets:* First, check your feedback on Gradescope; if you still have questions, contact your tutor.
- *I have a question about the marking of my assessed work on the R worksheets:* You can email me about this.
- *Due to truly exceptional and unforeseeable personal circumstances I require an extension on or exemption from assessed work:* You can apply by filling in the mitigating circumstances form at this link. Neither I nor your tutor can unilaterally offer an extension or exemption, so please don't ask. (Only exemptions, not extensions, are available for R worksheets.)

## Content of MATH1710

### Prerequisites

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

### Syllabus

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

An outline plan of the topics covered is the following.

- **Exploratory data analysis** [2 lectures]: Summary statistics, data visualisation
- **Probability** [16 lectures]:
  - Probability with events: Probability spaces, probability axioms, examples and properties of probability, “classical probability” of equally likely events, independence, conditional probability, Bayes’ theorem [6 lectures]
  - Probability with random variables: Discrete random variables, expectation and variance, binomial distribution, geometric distribution, Poisson distribution, multiple random variables, law of large numbers, continuous random variables, exponential distribution, normal distribution, central limit theorem [10 lectures]
- **Bayesian statistics** [2 lectures]: Bayesian framework, Beta prior, normal–normal model
- Summary and revision [2 lectures]

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

## Books

You can do well on this module by reading the notes and watching the videos, attending the lectures and tutorials, and working on the problem sheets and R worksheets, without needing to do any further reading beyond this. However, students can benefit from optional extra background reading or an alternative view on the material, especially in the parts of the module on probability. These books are also a good place to look if you want extra exercises and problems for revision.

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

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

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

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

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

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

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

For R, there are many excellent resources online.

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

## About these notes

These notes were written by Matthew Aldridge in 2021, and were edited and updated in 2022. They are based in part on previous notes by Dr Robert G Aykroyd and Prof Wally Gilks. Dr Jason Susanna Anquandah and Dr Aykroyd

advised on the R worksheets. Dr Aykroyd's help and advice on many aspects of the module was particularly valuable.

These notes (in the web format) should be accessible by screenreaders. If you have accessibility difficulties with these notes, contact me.

# Part I: Exploratory data analysis





# Chapter 1

## Summary statistics

### 1.1 What is EDA?

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

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

- **Summary statistics:** That is, calculating numbers that briefly summarise the data. A summary statistic might tell us what “central” or “typical” values of the data are, how spread out the data is, or about the relationship between two different variables.
- **Data visualisation:** Drawing a picture based on the data is 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 next to a double hashsign **##**. The **[1]** can be ignored (this is just R's way of saying that this is the first part of the answer – but the answer here only has one part anyway). Plots produced by R are displayed in these notes as pictures.

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

## 1.3 Statistics of centrality

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

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

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

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

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

Some measures of centrality are the following.

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

- The **mode** is the most common value of  $x_i$ . (If there are multiple joint-most common values, they are all modes.)
- Suppose the data is ordered as  $x_1 \leq x_2 \leq \dots \leq x_n$ . 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

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

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

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

59, 59, 59, 59, 60, 60, 60, 61, 62, 62, 62, 63, 63.

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

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

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  $\alpha$ -**quantile**  $q(\alpha)$  of the data is the datapoint  $\alpha$  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

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

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

The formula we've given for sample variance is sometimes called the “definitional formula”, as it's the formula used to *define* the sample variance. We can

rearrange that formula as follows:

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

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

This has left us with

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

This is sometimes called the “computational formula”; this is because it usually takes fewer presses of calculator buttons to compute the sample variance with this formula rather than the definitional formula. (But make sure you keep enough decimal points in  $\bar{x}^2$ .)

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

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

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

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

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

## Summary

- Exploratory data analysis is about taking a first look at data.

- Summary statistics are numbers calculated from data that give us useful information about the data.
- Summary statistics that measure the centre of the data include the mode, median, and mean.
- Summary statistics that measure the spread of the data include the number of distinct outcomes, the interquartile range, and the sample variance.

## Chapter 2

# Data visualisations

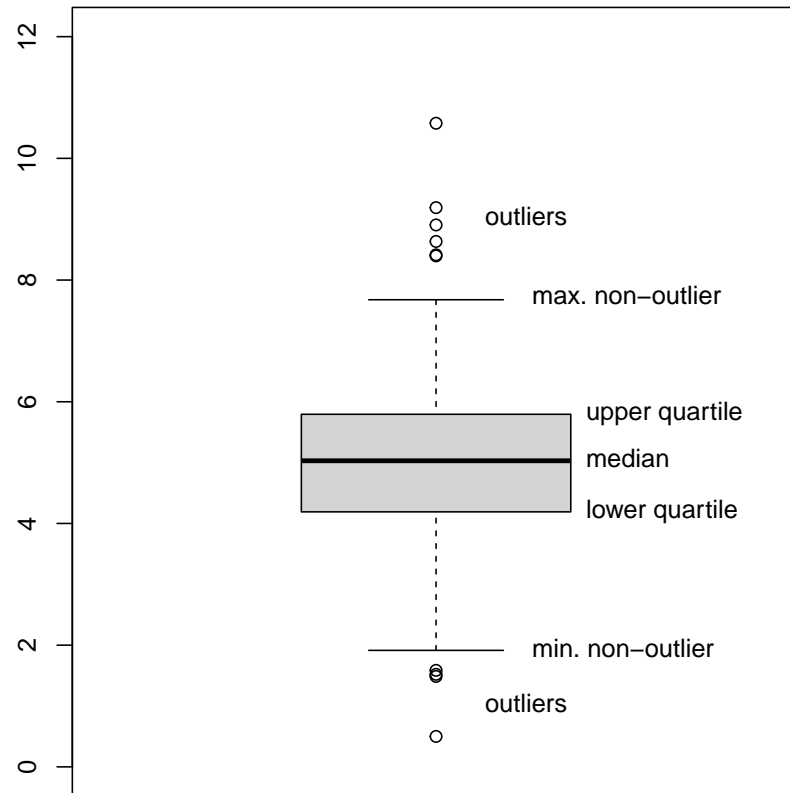
Data visualisations – drawings or graphs based on data – can help us to understand the “shape” of a dataset as part of exploratory data analysis. In this lecture, we’ll look at three types of data visualisation.

### 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
<b>Total</b>	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
<b>Total</b>	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 than the “15–30 minutes” bin, but they’re squashed into a much smaller bin.

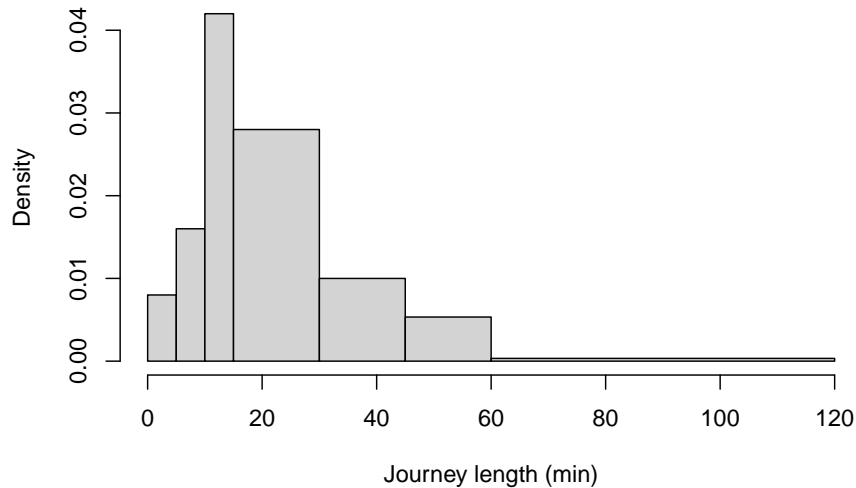
Data in bins can be illustrated with a **histogram**. A histogram has the measurement on the x-axis, with one bar across the width of each bin, where bars are drawn up to the height of the corresponding frequency density. Note that this means that the area of the bar is exactly the relative frequency 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:

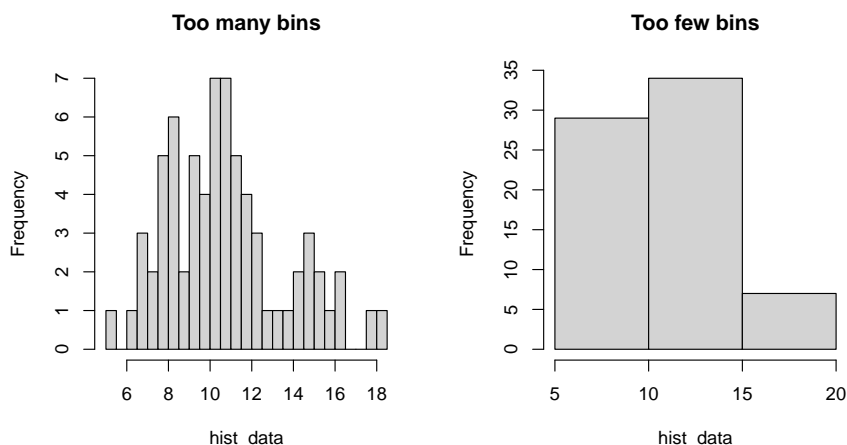
```
journeys <- read.csv("https://mpaldrige.github.io/math1710/data/journeys.csv")
bins <- c(0, 5, 10, 15, 30, 45, 60, 120)

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



Often we draw histograms because the data was collected in bins 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).

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



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

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

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

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

## 2.3 Scatterplots

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

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

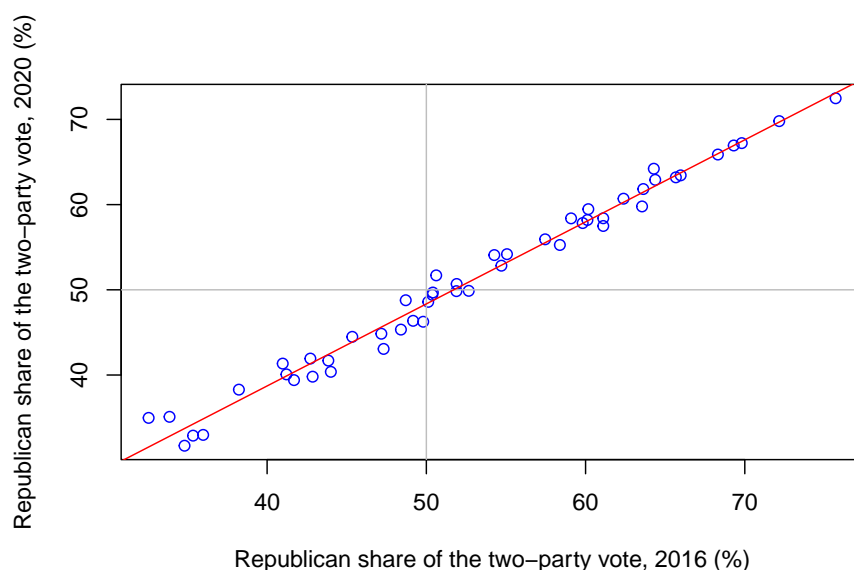
A good way to visualise the relationship between two variables is to use a **scatterplot**. In a scatterplot, the  $i$ th data pair  $(x_i, y_i)$  is illustrated with a mark (such as a circle or cross) whose x-coordinate has the value  $x_i$  and whose y-coordinate has the value  $y_i$ .

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://mpaldrige.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**

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

and  $s_x = \sqrt{s_x^2}$  and  $s_y = \sqrt{s_y^2}$  are the standard deviations.

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

straight line with a downward slope (big  $x$  = small  $y$ ); and values of  $r_{xy}$  near 0 indicate that there is a weak linear relationship between  $x$  and  $y$ .

For the elections data, the correlation is

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

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

which, as we expected, is extremely high.

## Summary

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

# Problem Sheet 1

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

## A: Short questions

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

:::: {myq} **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.

:::: {subq} **(a)** Calculate the mean number of Skittles in each packet. ::: {myanswers data-latex="" } *Solution.* This was in the notes:

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

:: :::: {subq} **(b)** Calculate the sample variance using the computational formula. ::: {myanswers data-latex="" } *Solution.*

$$\begin{aligned} s_x^2 &= \frac{1}{13-1} ((59^2 + 59^2 + \dots + 63^2) - 13 \times 60.6923^2) \\ &= \frac{1}{12}(47915 - 47886.2) \\ &= 2.40 \end{aligned}$$

**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  $\bar{x}$ . `:: :::: {subq}` **(c)** Calculate the sample variance using the definitional formula. `:: {myanswers data-latex=""}` *Solution.*

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

`:: :::: {subq}` **(d)** Out of (b) and (c), which calculation did you find easier, and why? `:: {myanswers data-latex=""}` *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. `:: :::: :::: {myq}` **A2.** Consider the following data sets of the age of elected politicians on a local council. (The "18–30" bin, for example, means from one's 18th birthday to the moment before one's 30th birthday, so lasts 12 years.)

Age (years)	Frequency	Relative frequency	Frequency density
18–30	1		
30–40	3		
40–45	4		
45–50	5		
50–55	3		
55–60	1		
60–70	3		
<b>Total</b>	20	1	—

`:: :::: {subq}` **(a)** Complete the table by filling in the relative frequency and frequency densities. `:: {myanswers data-latex=""}` *Solution.*

Age (years)	Frequency	Relative frequency	Frequency density
18–30	1	0.05	0.0041
30–40	3	0.15	0.015
40–45	4	0.2	0.04
45–50	5	0.25	0.05
50–55	3	0.15	0.03
55–60	1	0.05	0.01
60–70	3	0.15	0.015
<b>Total</b>	20	1	—

`:: :::: {subq}` **(b)** What is the median age bin? `:: {myanswers data-latex=""}` *Solution.* The 10th- and 11th-largest observations are both in the 45–50 bin, which is therefore the median bin. `:: :::: {subq}` **(c)** What is the modal age bin? `:: {myanswers data-latex=""}` *Solution.* The bin with the largest frequency density is 45–50, which is therefore the modal bin. `:: :::: {subq}` **(c)** Calculate (the standard approximation of) the mean age of the politicians. `:: {myanswers data-latex=""}` *Solution.* Pretending that each person is in the centre of their bin, we have

$$\bar{x} = \frac{1}{20} (1 \times 24 + 3 \times 35 + \dots + 3 \times 65) = \frac{946.5}{20} = 47.3.$$

`:: :::: {myq}` **3.** Consider the two datasets illustrated by the boxplots below. Write down some differences between the two datasets. `{r two-boxplots} set.seed(6) x <- rnorm(80, 10, 2) y <- 8 + rexp(80, 0.4) boxplot(x, y, names = c("Dataset 1", "Dataset 2")) ::::`



## B: Long questions

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

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

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

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

(b) Shirt sizes for a university football squad:

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

**B2.** A summary statistic is informally said to be “robust” if it typically doesn’t change much if a small number of outliers are introduced to a large dataset, or “sensitive” if it often changes a lot when a small number of outliers are introduced. Briefly discuss the robustness or sensitivity of the following summary statistics: (a) mode; (b) median; (c) mean; (d) number of distinct outcomes; (e) inter-quartile range; and (f) sample variance.

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

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

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

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

**B4.** A researcher wishes to study the effect of mental health on academic achievement. The researcher will collect data on the mental health of a cohort of students by asking them to fill in a questionnaire, and will measure academic

achievement via the students' scores on their university exams. Discuss some of the ethical issues associated with the collection, storage, and analysis of this data, and with the publication of the results of the analysis. Are there ways to mitigate these issues?

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

## C: Assessed questions

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

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

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

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

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

1.306, 1.301, 1.290, 1.266, 1.268, 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) Prove the following computational formula for the sample covariance:

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

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

### Solutions to short questions

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



# R Worksheets

## R worksheets

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

Week	Worksheet	Deadline for assessed work
1	<b>R basics</b> (Solutions)	—
2	Vectors	—
3	Data in R	Monday 24 October (Week 4)
4	Plots I: Making plots	—
5	Plots II: Making plots better	Monday 7 November (Week 6)
6	RMarkdown (optional)	—
7	Discrete distributions	Monday 21 November (Week 8)
8	Discrete random variables	—
9	Normal distribution	Monday 5 December (Week 10)
10	Law of large numbers	—
11	Recap	Thursday 15 December (Week 11)

## About R and RStudio

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

R and RStudio are free/open-source software.

## How to access R and RStudio

There are a few ways you can access R and RStudio.

First, you can **install R and RStudio on your own computer**. Students who have their own computer (with administration and installation rights) usually find this the most convenient way use R.

When you install R and RStudio, it's important that you install R (the programming language) first, and only install RStudio (the program to use R) after R has already been installed. This ensures that RStudio can “find” R on your computer.

1. *First*, install R. Go to the Comprehensive R Archive Network (CRAN) and follow the instructions:
  - Windows: Click “Download R for Windows”, then “Install R for the first time”. The main link at the top should be to download the most recent version of R.
  - Mac: Click Download R for macOS, and then download the relevant PKG file. (For typically older Intel-based Macbooks, you must use the “Intel 64-bit build”; for post-November 2020 M1 or M2-based “Apple silicon” Macbooks, the “Apple silicon arm64 build” may be slightly faster.)
2. *After* R is installed, *then* install RStudio. Go to the Download page at [RStudio.com](https://www.rstudio.com) and follow the instructions. You want “RStudio Desktop”, and you want the free version.

If you have difficulty installing R, come along to the R troubleshooting drop-in session in Week 2 and bring your computer with you (if it's sufficiently portable), and we'll do our best to help.

Second, you can **use R and RStudio on University computers**. All University computers have access to R and RStudio, via the AppsAnywhere service. Again, you should first install R (which, at the time of writing, is confusingly included under the name “CRAN”) via AppsAnywhere, and then install RStudio via AppsAnywhere.

The R drop-in sessions take place in computer rooms, so if you have problems accessing R and RStudio on University computers, you can get help at the drop-in sessions too.

Third, you can **use the RStudio Cloud**. The RStudio Cloud is a cloud-hosted “Google Docs for R” that you can use through your web browser – you can get 25 hours per month for free, which should be plenty for this module, or pay for more.

If you have access to a computer on which you can't install software, such as some Chromebooks or tablet computers, or if you're borrowing a friend's laptop, the RStudio Cloud can be a convenient solution.

## R troubleshooting drop-in sessions

You will learn to use R by working through the R Worksheets. Learning to use a programming language is different from learning mathematics: you should expect to regularly get frustrated and annoyed when the computer seems to refuse to do what you want it to (but also occasionally experience the joy of getting it right!). This is a normal part of learning.

However, many students find getting with started with R in the first few weeks particularly difficult. Also, sometimes students have problems installing R and RStudio on their own computers. To help with this, we have organised optional R troubleshooting drop-in sessions in Weeks 2 and 3. Check your timetable for details – they are probably listed as “computer practicals”.