



Newcastle University

MAS2905

Probability and Bayesian Methods with R

Part 1

Dr. Lee Fawcett

Semester 1, 2020/21

1 General housekeeping

1.1 Lecturer information

The module leader for MAS2905, and lecturer for both semesters, is Dr. Lee Fawcett. If you have any questions about the module he can be contacted via email at `lee.fawcett@newcastle.ac.uk`. Lee is based in room 2.07 on the second floor of the Herschel Building.

There will be several other people assisting with this module in semester 1, primarily in the computer practical sessions. They're all really friendly PhD students – **Kieren Peel**, **Amy Green**, **Tom Lowe** and **Sophie Harbisher** – give them a shout in class if you're stuck with anything!

Photos to be inserted here.

1.2 Classes

As you probably know, things will be very different in 2020/21. You will have a blended learning experience, meaning you will have some *Present-in-Person* activities on campus and some online activities. Your online activities might be *synchronous* or *non-synchronous*.

For semester 1 of MAS2905, this is how we will use these activities:

- Present-in-Person (PiP): A series of 90 minute supervised computer practical sessions in the Herschel PC cluster. These take place roughly every three weeks, and will appear on your timetable and will be advertised on Canvas (and are detailed in Section 1.3)
- Non-synchronous online (NSO): Online activities posted to Canvas every week, covering what would have ordinarily been covered in standard face-to-face lectures. This will mainly consist of directed reading (e.g. lecture notes), completing exercises/practice questions, summaries/discussions of learning objectives, assessment feedback and other 'fun' activities (e.g. quizzes!). These activities must be completed in your own time, whenever you can fit them in, but it is *vital* that they are completed by their completion dates as advertised on Canvas (and as shown in Section 1.3)
- Synchronous online (SO): Timetabled sessions, broadcast live online **via Zoom**. No new material will be presented – instead, we will work through problems and exercises based on the theory you have studied in the NSO activities. Some SO sessions will be dedicated 'office hours' held via Zoom, for those who want to drop-in (e.g. if you're stuck on an assignment question). You should make every effort to attend these SO when they appear on your timetable; the dates/times of these sessions are also advertised on Canvas (and shown in Section 1.3)

Students can take part in other NSO activities, should they wish, including interacting with the Canvas discussion board; optional Zoom drop-in sessions; and one-to-one Zoom meetings with the module leader (arranged via email in advance).

And of course, you will all be expected to prepare and complete homeworks in your own time, and use guided independent reading to help prepare for the tests!

Note that this arrangement is specific to semester 1 of this module, where the primary focus will be on statistical computing and R. Our PiP sessions in semester 2 will be classroom-based, as the primary focus will be on Bayesian methods. The specific arrangements for semester 2 will be discussed in the booklet for Part 2 of this module, distributed in January 2021.

1.3 Schedule

A detailed week-by-week schedule to appear here!

1.4 Assessment

The full 20 credit MAS2905 module will be assessed via a formal written examination in summer 2021 (likely a '24 hour take-home paper') and some in-course assessment. The make-up is 50% summer exam, 50% in-course assessment. Further details:

- Summer exam, 100 marks in total
 - Semester 1 material: Section A (20 marks)
 - Semester 2 material: Section B (80 marks)
 - Answer *all* questions in *both* parts
- In-course assessment, 100 marks in total
 - Semester 1 material: One homework (20 marks), two tests (30 marks each)
 - Semester 2 material: Two short homeworks (10 marks each)

Exam questions in section A, relating to the material from semester 1 of the module, will focus mainly on the theory and probability/statistics developed throughout the semester (rather than programming in R). The semester 1 homework will be R-based, and the semester 1 tests will require the use of R to solve some problems in probability/statistics. A sample test (with solutions) will be made available via Canvas. **More details about the tests to go here**

More details about the semester 2 assessment will be discussed in the booklet for Part 2 of this module, distributed in January 2021.

1.5 Late work policy

Late work policy to be inserted here.

1.6 Course materials

All course materials will be posted to the dedicated Canvas page for MAS2905. Here, you will find:

- A full copy of these lecture notes, with gaps
- A copy of the lecture notes with the gaps filled in, after the associated PiP/SO/NSO activities
- All slides used to support PiP/SO/NSO activities have taken place
- All ReCap recordings
- Solutions to assessed exercises
- Any other materials, including additional handouts/worksheets, links to useful Numbas material, specimen exam papers and solutions, . . .

You should have a hard copy of these lecture notes before the start of the course – these will be available to collect from the General Office in the Herschel Building.

1.7 Books and other resources

At various points in the module you will be directed to additional reading, should you want extra examples or just another source of information relating to the material presented in the notes. This will primarily be text books (many of which are available to view online), but perhaps other sources (e.g. journal publications, magazine articles etc.).

Some useful books for semester 1 include:

- W.N. Venables and D.M. Smith: An Introduction To R. Free to download at:

<https://cran.r-project.org/doc/manuals/R-intro.pdf>

- D. Braun and D.J. Murdoch: A First Course In Statistical Programming With R
- S. Ross: A First Course in Probability
- N.A. Weiss, P.T. Holmes and H. Michael: A Course In Probability

More details about semester 2 reading material will be provided in the booklet for Part 2 of this module, distributed in January 2021.

1.8 General background for Part 1 of MAS2905

In Part 1 of the module we will learn about computational tools for performing probability calculations and statistical analysis. We will use the R programming language. R is very widely used in universities, businesses and other organizations due to its many advantages over other similar software packages. It is available free of charge, it is 'open source' and can be used for a very wide range of statistical and scientific applications. Libraries of R functions and data-structures have been created and are readily available on the web. These enable statisticians to perform a very wide range of tasks but with a minimum of additional programming. Learning to use R is important because:

- (i) It is the foremost tool in modern computational statistics
- (ii) It is a great way to learn general concepts in programming which can be used with other languages or in other contexts
- (iii) It can be used to illustrate ideas in probability and statistics

In MAS1802 last year we focused mainly on point (ii) above; in this module, we focus mainly on points (i) and (iii).

Given the COVID-19 situation in 2020 (and the industrial action that took place immediately before the closure of our campus), many of you will be worried about not having had the chance to learn R properly last year, in MAS1802. Thus, one of the first things we will do in MAS2905 is review some of the key concepts in the MAS1802 syllabus. This will roughly occupy the first two weeks of the module.

For the rest of semester 1, we will present a mixture of background theory in probability and statistics, together with elements of the R programming language. Thus, from around the third week of the course, activities will contain both some statistical theory and some R programming concepts. Most of the theory will be covered in your own time, during the non-synchronous online (NSO) activities that you are set in Canvas. In our synchronous online (SO) and Present-in-Person (PiP) sessions, we will put that theory into practice through worked examples, both written and on the computer in R.

All materials for semester 1 of the module are in this booklet, including:

- Background theory that you will be expected to read yourself through NSO activities
- Pencil-and-paper exercises – some of which you will try out yourself in NSO activities, and some of which we will work through together in SO sessions
- Computer practical exercises – some of which you will try out yourself in NSO activities, and some of which we will work through together in our PiP sessions in the computer cluster

It is vital that, if possible, you attend your PiP sessions in the computer cluster. This is a great opportunity to ask us questions face-to-face, seek help from other students, and of course prepare

for the two tests. It is even more important that you follow instructions on Canvas and complete the other activities by the dates suggested – if you are told that some NSO activities should be completed before the next PiP computer practical, then make sure you do exactly that – the work set in the computer practical class will assume you have completed all work to-date!

2 R Refresher

2.1 Preliminaries

R is installed on all University machines. In this class we will be using RStudio, which is an R *Integrated Development Environment*, or IDE. This just makes R much easier to use and navigate, through the four windows that you see when you open RStudio (text editor, console window, file explorer and graphics display).

You might remember from MAS1802 that both R and Rstudio are free, so you can install them on your own computer. Versions of R and RStudio are available for Windows, Apple Mac and Linux. See

<http://www.ncl.ac.uk/maths-physics/students/resources/>

for more details.

We will assume a limited working knowledge of R and RStudio. The first two weeks are designed to recap the basics from MAS1802. Specifically, we will review (i) dataframe manipulation; (ii) graphics; (iii) functions and control statements. Everything we do for the remainder of semester 1 will then assume you have a good grasp of these basics, so please work hard to make sure you're up-to-speed with everything as we progress through this Section of the notes! And of course, get in touch if you're not sure about anything – either by private email or via the Canvas discussion board.

You might recall the `movies` dataset from MAS1802. This was derived from the internet movie database (IMDB): <http://imdb.com/>, which is a website devoted to collecting movie data supplied by studios and fans. It claims to be the biggest movie database on the web and is run by *Amazon*. More information can be found online:

http://imdb.com/help/show_leaf?about

including information about the data collection process. This is a great dataset to use to demonstrate R's dataframe manipulation and querying capabilities – it is large, has many variables, and is interesting. And everyone loves movies! See Table 1 for an extract of the movies dataset.

Title	Year	Length	Budget	Voting statistics					Movie genre							
				Rating	Votes	r1	...	r10	mpaa	Action	Animation	Comedy	Drama	Documentary	Romance	Short
A.k.a. Cassius	1970	85	NA	5.7	43	4.5	...	14.5	PG	0	0	0	0	1	0	0
AKA	2002	123	NA	6.0	335	24.5	...	1.5	R	0	0	0	1	0	0	0
Alien Vs. Pred	2004	102	45000000	5.4	14651	4.5	...	4.5	PG-13	1	0	0	0	0	0	0
Abandon	2002	99	25000000	4.7	2364	4.5	...	4.5	PG-13	0	0	0	1	0	0	0
Abendland	1999	146	NA	5.0	46	14.5	...	24.5	R	0	0	0	0	0	0	0
Aberration	1997	93	NA	4.8	149	14.5	...	4.5	R	0	0	0	0	0	0	0
Abilene	1999	104	NA	4.9	42	0.0	...	24.5	PG	0	0	0	1	0	0	0
Ablaze	2001	97	NA	3.6	98	24.5	...	14.5	R	1	0	0	1	0	0	0
Abominable Dr	1971	94	NA	6.7	1547	4.5	...	14.5	PG-13	0	0	0	0	0	0	0
About Adam	2000	105	NA	6.4	1303	4.5	...	4.5	R	0	0	1	0	0	1	0

Table 1: The first ten rows of the movie data set. **Credit:** This data set was initially constructed by Hadley Wickham at <http://had.co.nz/>.

We have written an add-on R package for the module which will allow easy access to many interesting datasets (including the movies dataset). Installing the package is straightforward, and will make loading datasets used in the notes really easy:

```
1 install.packages("ncldata", repos = "http://R-Forge.R-project.org",  
2 type = "source")
```

To load the package, type:

```
1 library(ncldata)
```

Then, to specifically load the movies dataset, type:

```
1 data(movies)
```

Alternatively, some datasets can be downloaded directly from the course Canvas page; examples of how to do this will be given later in the notes.

2.2 Data types in R