

Instructions

You should attempt **both** parts of this project. Solutions should be typed into a Microsoft Word document and any relevant plots/code copied-and-pasted from RStudio into this document. You will find it helpful if you have your lecture notes to hand while you work through this project – if you no longer have a copy of these, they can be found on the MAS1802 Blackboard page or on the dedicated webpage:

www.mas.ncl.ac.uk/~nlf8/teaching/mas1802/

Part 1: Plotting and Summaries

In this part of the project you will demonstrate your ability to use R's plotting and data summary functions. It will help if you have your lecture notes for **Chapter 2** in front of you. You may also be required to manipulate objects in R, and so **Chapter 1** of the lecture notes may also be useful. Throughout this part you will be analysing a unique subset of the Movies database used throughout the course. You should use the following code to load your unique database into R, where `LOGIN_ID` is your computer ID without the 'a' or 'b' at the start (**type extremely carefully but do not copy-and-paste the code; notice that the third line of code has wrapped onto the next line – you should enter this as a single line of code**):

```
set.seed (LOGIN_ID)
a = sample(100:2500, 1)
movies = read.table("http://www.mas.ncl.ac.uk/~nlf8/movies.txt",
  skip = a, header = T)
attach (movies)
```

You must provide solutions to all questions below. Marks will be awarded for style and presentation – think carefully about labels on axes and the titles of any plots produced!

1. How many people submitted votes for the films "*Alien versus Predator*", "*Midnight in the Garden of Good and Evil*" and "*Zoolander*"? What is the mean number of votes per film across all films in your database? What is the standard deviation? **[5 marks]**
2. How many films are classified as "Action"? How many are classified as "Romance"? How many "Short" films are there in your database? **[3 marks]**
3. Produce two box and whisker plots: One for movie budgets and one for movie budgets split by MPAA rating. Include both boxplots in your solutions, and comment. *Make sure you include only those films for which the budget was known!* **[10 marks]**
4. For films made between the years 1970 and 2000 (inclusive),
 - (a) produce a bar chart to show the number of films made by year, and comment; **[5 marks]**
 - (b) find the median and interquartile range for the movie lengths; **[5 marks]**
 - (c) use Scott's rule to produce a histogram of movie lengths, coloured 'forest green', and comment on the shape of the distribution. On this histogram superimpose a vertical red line at the median movie length. **[10 marks]**
5. Produce a plot of mean movie rating conditional on its budget (where the movie budget was actually known). Include this plot in your solutions, and comment. **[12 marks]**

Part 2: Monte Carlo Integration

In this part of the Project you will use Monte Carlo techniques to estimate the area under a curve – i.e. you will perform *Monte Carlo Integration*. It will help if you have your lecture notes for [Chapter 6](#) in front of you – in particular, [Section 6.3](#) (starting on page 44). **You must provide solutions to all parts of questions 1 and 2.**

1. In this question, you are required to use Monte Carlo Integration to approximate $\int_0^1 f(x)dx$, where

$$f(x) = abx^{a-1}(1-x^a)^{b-1}$$

for $0 \leq x \leq 1$ and where $a > 0$ and $b > 0$.

- (a) Each student will need to generate their own unique values for a and b . To generate *your* values for a and b , type the following code into R:

```
set.seed (LOGIN_ID)
a = sample (3:14, 1); b = sample (3:14, 1)
```

where `LOGIN_ID` is your computer ID without the 'a' or 'b' at the start. In your solutions, write down your values for a and b and the resulting integrand. For example, if your generated values for a and b were 10 and 12 respectively, you would write down:

$$a = 10; \quad b = 12; \quad f(x) = 120x^9(1-x^{10})^{11}.$$

- (b) Use R to produce a plot of $f(x)$. Within the `plot` command, use the argument `xlim = c(-1, 2)`. Use the `abline` command to indicate a suitable *simulation grid* on your plot. Include this plot in your solutions, *with appropriately labelled axes and title!* **[7 marks]**
- (c) Using a `for` loop over $N = 10^6$ replications, estimate your integral. Include your R code in your solutions, which should be nicely formatted/indented. Where necessary, include comments in your code so it is clear what you are trying to do. In your solutions, make it obvious what your estimated integral is. **[18 marks]**

2. A special function that comes up a lot in Statistics is the probability density function (p.d.f.) of the standard Normal distribution, that is, the p.d.f. for Z , where $Z \sim N(0, 1)$; see your MAS1604 lecture material. The p.d.f. for Z is a bell-shaped curve centred around zero, given by:

$$g(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right), \quad -\infty < z < \infty.$$

Using Monte Carlo Integration estimate $\int_{-d}^d g(z)dz$, where d is the value:

```
set.seed (LOGIN_ID)
d = round (runif (1, 0.5, 2.5), 2)
```

Include in your solutions exactly what you did for question 1: your value for d ; a plot of the function over a suitable range, showing the simulation grid; your nicely formatted/indented R code used to perform the Monte Carlo Integration, appropriately commented; your final answer (i.e. the estimate of your integral). You should also include vertical lines at the points $\pm d$ on your plot. **[25 marks]**