1 R Refresher

Querying databases; plots and summaries

You might remember the movies database from MAS1802. This database is available from the MAS2602 webpage – you can load it into R by typing the following code:

```
1 movies = read.table("http://www.mas.ncl.ac.uk/~nlf8/movies.txt", header=T)
2 attach(movies)
```

Listing 1: Loading the movies database

(a) Use the `head` command to look at the first six rows of the database. How many variables are there in the dataset?

(b) Use the `dim` command to find out the dimensions of the database. How many movies are there?

(c) How many people submitted votes for the films *"AVP: Alien Vs. Predator"*, *"Midnight in the Garden of Good and Evil"* and *"Zoolander"*? (Hint: the `which` and `tail` commands might come in useful here)

(d) What is the mean number of votes per film across all films in the database? What is the standard deviation?

(e) How many films are classified as "Action"? How many are classified as "Romance"? How many "Short" films are there in the database? (Hint: use the `sum` command)

(f) Type the following code:

```
1 Budgets_known = Budget[Budget!=(-1)]
2 par(mfrow=c(2,2))
3 hist(Budgets_known)
4 abline(v = mean(Budgets_known), col = "red")
5 boxplot(Rating, main = "Boxplot for movie ratings")
6 barplot(table(mpaa), xlab = "MPAA rating", col = "misty rose")
7 apply(movies[ ,3:6], 2, mean)
8 Length2 = Length[Year>=1970 & Year<=2000]
```

Listing 2: Querying the movies database

Using the code in Listing 2:

(i) Can you figure out what is going on in line 1?

(ii) What is the purpose of line 2?

(iii) Re-produce the plots, but this time colour the histogram "forest green", change its title, and draw a vertical line at the median budget; colour the boxplot "wheat"; and add the word "frequency" to the y-axis of the bar chart.

(iv) Produce a scatterplot of Rating against Votes, and explain what you see.

(v) What is going on in line 7?

(vi) Find the median and interquartile range of movie lengths, for movies made between the years 1970 and 2000 (inclusive) (Hint: the command `quantile` might be useful here)

(vii) Write a line of code that will extract the movie names, and the years in which the movies were made, into a new dataframe - and store this new dataframe in `movies2`.

Functions, for loops and if statements

Suppose the random variable Z follows a standard Normal distribution, that is, $Z \sim N(0,1)$. Then Z has probability density function given by

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \quad z \in \mathbb{R}.$$

Type the following code into R:

```
z = seq(-4, 4, 0.001)
d = 1/(sqrt(2*pi))*exp(-(z^2)/2)
plot(z, d, xlab = "z", ylab = "density", type = "l")
abline(v = -2, lty = 2, col = "red")
abline(v = 1, lty = 2, col = "red")
abline(h = 0.4, lty = 2, col = "red")
```

Listing 3: Drawing the standard Normal distribution with simulation grid

We will now use Monte Carlo integration to estimate $P(-2 < Z < 1)$ by following the steps below:

1. Throw randomly generated points at a "sampling grid" that covers the region of interest (see red dashed lines on your plot)
2. Calculate the proportion of points P that lie under the density curve for Z
3. Find
$$P(-2 < Z < 1) = \int_{-2}^{1} f_Z(z)dz \approx P \times \text{Area of sampling grid},$$
i.e. the area under the Normal curve and between the vertical red lines in your plot.

   (Hint: look back at your MAS1802 notes, chapter 6, for more information on this).

Type in the following code:

```
MCintegrate = function(N){
  no_of_hits = 0
  for(i in 1:N){
    x = runif(1, -2, 1)              #Generate the x co-ordinate
    y = runif(1, 0, 0.4)             #Generate the y co-ordinate
    d = 1/(sqrt(2*pi))*exp(-(x^2)/2) #Evaluate the Normal density at x
    if(y<d){                         #Is the point under the curve?
      no_of_hits = no_of_hits+1      #If so, increase the counter
    }
  }
  P = no_of_hits/N                   #Find P
  area_under_curve = P*(3*0.4)       #Find the area under the curve
  return(area_under_curve)}
```

Listing 4: Using Monte Carlo integration to find probabilities from the standard Normal distribution

(a) Run the function in Listing 4 using `N` = 10, 100, 1000 and 10,000.

(b) Can you generalise the function in Listing 4 so that the function can be used to find probabilities from the standard Normal distribution between *any* two points (as opposed to just between –2 and 1)?

2 Simulation from probability distributions

1. Chapter 1 of your lecture notes will help you with the following questions.

   (a) Use R to generate a sample of 100 observations from W, where $W \sim Po(2.5)$. Produce a barplot of this sample.

   (b) We are interested in X, the number of broken eggs in a box of 12 eggs. Suppose there is a 1% chance that an egg is broken. Use R to find:

   (i) the probability that a box has exactly three broken eggs;

   (ii) the probability that a box has fewer than three broken eggs.

   (iii) Generate 1000 values from X and use your simulated values to verify the mean and variance of X.

   (c) Suppose that Y is used to represent the number of minutes late a bus arrives at the station. It is assumed that $Y \sim U(-15, 45)$.

   (i) Do you think this is a reasonable model to use in this scenario? Explain your answer.

   (ii) In R, generate a sample of 1000 observations from Y, and use your sample to verify that

   $$E[Y] = (a+b)/2 \quad \text{and} \quad Var(Y) = \frac{(b-a)^2}{12}.$$

   (iii) Produce a histogram of your sample, with appropiately labelled axes.

2. Daily maximum wind speeds, taken from hourly records, are assumed to follow an exponential distribution. Data for 20 consecutive days at a location in the USA are recorded below (miles per hour):

26	10	10	3	5	33	13	23	5	8	14	24	3	2	7	10	46	15	2	35

   (a) Produce a histogram of these wind speed maxima. Within the `hist` command, use the argument `freq = FALSE`; label your axes appropriately, and give your histogram a suitable title; and make the x-axis extend from 0 to 80 by using `xlim = c(0, 80)` within the `hist` command.

   (b) You should know from MAS2901 that the maximum likelihood estimate (MLE) for the rate parameter λ in the exponential distribution is

   $$\hat{\lambda} = \frac{1}{\bar{x}}.$$

   Use R to find the MLE for λ using the wind speed data observed at this location, and store this value in `lambda`.

   (c) The exponential distribution has probability density function (PDF) given by

   $$f_X(x) = \lambda e^{-\lambda x}, \quad \lambda > 0.$$

   Use the following code to superimpose your fitted PDF on the histogram:

```
x = seq(0, 80, 0.001)
d = lambda*exp(-lambda*x)
lines(x, d, type="l", col="red")
```

Listing 5: Overlaying fitted exponential PDF

(d) Does your fitted exponential distribution provide a good fit to the wind speed data?

(e) During a hurricane, observed wind speeds exceed 70 miles per hour. Use R to find the probability, according to your fitted model, of wind speeds exceeding 70 miles per hour.

(f) How does your calculation in part (e) change if a new data point of 72 miles per hour is included in the analysis?

3. Use the `pnorm` command in R to confirm your Monte Carlo integration approximation to $P(-2 < Z < 1)$, $Z \sim N(0, 1)$, from the R Refresher part of this practical sheet.

4. Write an R function to generate samples from a 3-component normal mixture distribution with parameters

$$\begin{array}{lll} \mu_1 = 2, & \sigma_1 = 1, & w_1 = 0.3, \\ \mu_2 = 5, & \sigma_2 = 2, & w_2 = 0.4, \\ \mu_3 = 7, & \sigma_3 = 0.5, & w_3 = 0.3. \end{array}$$

Simulate a sample of size 1000 from this distribution and plot a histogram.