MAS115: R programming
Lecture 4: Monte-Carlo Estimation
Lab Class: more loops via `while` and `repeat`

Paul Blackwell

The University of Sheffield
School of Mathematics and Statistics

Aims

In lab we will learn two more methods of creating loops in **R**

- `while`
- `repeat`

Monte-Carlo methods:

- Formalisation
- Wind Farms
- Chemical Plant

Demonstrate how to use our pseudo-code to write our program step-by-step

Monte Carlo simulation

If X is a random variable, and we want to estimate the $\mathbb{E}[h(X)]$ for any function $h(\cdot)$, we can use

$$\mu = \mathbb{E}[h(X)] \approx \frac{1}{n}\sum_{i=1}^{n} h(X_i)$$

if X_i are a *large* sample from the density $f_X(\cdot)$ of X.

Monte Carlo simulation

In fact—as those of you taking MAS113 will see—it can be shown that, for large n,

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i) \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Our Monte Carlo estimate will be centred on the true value, with a variance around it that decreases with n in a well-understood way.

Monte Carlo simulation

An important special case is when $h(\cdot)$ is an *indicator function*, taking the value 1 when some event involving X occurs, and 0 otherwise.
Some times written $\mathbb{1}_E$ where E is the event.
For example, we can estimate the probability that $X < 0$ by taking $h(X) = \mathbb{1}_{\{X<0\}}$;

$$p = \Pr\{X < 0\} \approx \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i<0\}}.$$

Monte Carlo integration—the trick

Often, the value I of a definite integral can be written as the mean of a function of a random variable—even though I is not random. Then we can approximate the integral:

- get our computer to sample lots of values of that random variable;
- work out the sample mean.

Monte-Carlo Simulation: Pseudo-Code to Computer Code

Situating a wind farm

Problem:

A particular site is being considered for a wind farm. At that site, Y_t, the log of the wind speed in m/s on day t is known to depend upon the previous two days' winds:

$$Y_t = 0.6Y_{t-1} + 0.4Y_{t-2} + \varepsilon_t,$$

with $\varepsilon_t \sim N(0, 0.01)$. If $Y_1 = Y_2 = 1.5$, what is the probability that the wind speed $\exp(Y_t)$ will be below 4 m/s for more than 10 days in a 100 day period?

Situating a wind farm

Problem:

A particular site is being considered for a wind farm. At that site, Y_t, the log of the wind speed in m/s on day t is known to depend upon the previous two days' winds:

$$Y_t = 0.6\,Y_{t-1} + 0.4\,Y_{t-2} + \varepsilon_t,$$

with $\varepsilon_t \sim N(0, 0.01)$. If $Y_1 = Y_2 = 1.5$, what is the probability that the wind speed $\exp(Y_t)$ will be below 4 m/s for more than 10 days in a 100 day period?

Solution style - Monte Carlo estimation

- Create n potential series of 100 day series of wind speeds.
- Find out the proportion with more than 10 days of low wind.
- If n is large this will be a good estimate of probability.

Situating a wind farm - Single Realisation

Let's break this up one part at a time. First let's create a single hypothetical set of wind speeds for the 100 days.

Initial Task:

Write pseudo code which generates a vector Y which

- Records a hypothetical set of log-wind speeds for 100 days with starting values $Y_1 = Y_2 = 1.5$
- Finds if there have been more than 10 days of winds below 4 m/s. Store this as

$$E = \begin{cases} 1 & \text{if more than 10 days below 4 m/s} \\ 0 & \text{if 10 or fewer days below 4 m/s} \end{cases}$$

Single Realisation Wind-Farm Pseudo-code

CREATE Y as vector of length 100
Set $Y_1 = Y_2 = 1.5$

1. FOR $(t = 3, 4, \ldots, 100)$:
   - Sample ε_t from $N(0, 0.01)$
   - Set $Y_t \leftarrow 0.6Y_{t-1} + 0.4Y_{t-2} + \varepsilon_t$

   ENDFOR
2. Count number of elements of $\{Y_1, \ldots, Y_{100}\}$ less than $\log 4$:
   - Set $X_i \leftarrow \sum_{t=1}^{100} \mathbb{1}_{[Y_t < \log 4]}$
3. Determine if event E has occured for time series i:
   - Set $E \leftarrow \mathbb{1}_{[X_i > 10]}$

Situating a wind farm - Full solution

Now we can create a single realisation and see if it had low-wind speeds. We can embed this in another `FOR` loop to create n hypothetical sets of 100 day wind speeds and use Monte Carlo to estimate the probability.

Define E: the event that in 100 days the wind speed is below 4 m/s for more than 10 days. To estimate $P(E)$, generate lots of individual time series, and count proportion of series in which E occurs.

Pseudo code to solve the whole problem:

- FOR $(i = 1, 2, \ldots, N)$:
  1. Generate ith realisation of the time series process:
     CREATE Y as vector of length 100
     Set $Y_1 = Y_2 = 1.5$
     FOR $(t = 3, 4, \ldots, 100)$:
     - Sample ε_t from $N(0, 0.01)$
     - Set $Y_t \leftarrow 0.6Y_{t-1} + 0.4Y_{t-2} + \varepsilon_t$

     ENDFOR
  2. Count number of elements of $\{Y_1, \ldots, Y_{100}\}$ less than $\log 4$:
     - Set $X_i \leftarrow \sum_{t=1}^{100} \mathbb{1}_{[Y_t < \log 4]}$
  3. Determine if event E has occured for time series i:
     - Set $E_i \leftarrow \mathbb{1}_{[X_i > 10]}$

  ENDFOR
- Estimate $P(E)$ by $\frac{1}{N}\sum_{i=1}^{N} E_i$

Pseudo-code Task 1 - Death at a Chemical Plant

Problem:

A fluid dynamics model describes concentration of a pollutant at any point in a region following release from a point source,

$$C(y,z) = \frac{Q}{2\pi u_{10}\sigma_z\sigma_y} \exp\left[-\frac{1}{2}\left\{\frac{y^2}{\sigma_y^2} + \frac{(z-h)^2}{\sigma_z^2}\right\}\right], \qquad (1)$$

where the variables have the following meanings: C: air concentration of pollutant; Q: release rate; u_{10}: wind speed at 10m above ground; σ_y, σ_z: diffusion parameters in horizontal and vertical directions; h: release height; (y,z): coordinates along wind direction and above ground.

We are given $Q = 100$, $h = 50$m, but u, σ_z, σ_y are uncertain. This means that $C(y, z)$ is a random variable, dependent upon the inputs u, σ_z, σ_y. If

$$\log u_{10} \sim N(2, 0.1) \qquad \log \sigma_y^2 \sim N(10, 0.2) \qquad \log \sigma_z^2 \sim N(5, 0.05),$$

what is the 95th percentile of $C(100, 40)$?
More generally, what is the distribution of $C(100, 40)$?

Class Pseudo Code Task:

Write pseudo code which solves this problem using Monte Carlo methods.

Create a large sample of hypothetical values for $C(100, 40)$ by sampling values of the input parameters u, σ_z, σ_y.

Given the sample, we can find the e.g. 95% quantile of these computer-created C values (i.e. the value that only 5% of samples are larger than). No need to give details of this follow-up calculation.

You need to write pseudocode which you could give to anyone else and they would be able to understand and implement it!

Pseudo-code Task 1 Solution

1. INPUT y, z, N.
2. CREATE C as a vector of length N.
3. SET $Q \leftarrow 100, h \leftarrow 50$.
4. FOR $i = 1, 2, \ldots, N$:
   - 4.1 Sample a set of input values:
     - Sample $u_{10,i}$ from $\log N(2, 0.1)$
     - Sample $\sigma^2_{y,i}$ from $\log N(10, 0.2)$
     - Sample $\sigma^2_{z,i}$ from $\log N(5, 0.05)$
   - 4.2 Evaluate the model output C_i. Set

$$C_i \leftarrow \frac{Q}{2\pi u_{10,i}\sigma_{z,i}\sigma_{y,i}} \exp\left[-\frac{1}{2}\left\{\frac{y^2}{\sigma^2_{y,i}} + \frac{(z-h)^2}{\sigma^2_{z,i}}\right\}\right]$$

   ENDFOR
5. Return $C_1, C_2, \ldots, C_N$.

Conclusion - This week you should understand

- Different ways to perform loops in **R** to iterate statements.
- A more formal description of Monte Carlo simulation
- Using pseudo-code to write proper code

In the lab class today we will be learning more about the different loops and implementing more pseudo-code to solve the chemical plant problem.