

Practical 2: Probability Distributions in R

Main Aim: To recap **simulation** and **probability calculations** in R.

There are two activities to complete:

1. **Using R's built-in functions:** Directly related to the material in Chapter 3 of the course notes – two quick quizzes and two bigger questions to try
2. **The Inverse CDF Method:** Directly related to the material in Chapter 4 of the course notes – some directed reading, followed by a guided example and a more independent example

All exercises should get you up-to-speed with probability distributions in R. You can either follow the instructions on this handout, or use the associated slides.

Note that you should not attempt this practical unless you have completed all activities from weeks 3 and 4, including:

- Attend (or watch the recording of) the synchronous online lectures on Wednesdays 4th and 11th November
- Complete all work associated with Home Study videos 3, 4 and 5
- Complete all exercises in Chapters 3 and 4 of the course notes

Of course, you must also have access to RStudio. If you still don't have RStudio installed on your machine, visit our student resources webpage:

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

From this link, click on the link for our "Computing page", log-in using your University details, and then click on the link for "R". Here you will find some basic resources aimed at students who have never used R before; scroll down, and you'll find information about how to install R on your home PC. Alternatively, Chris Graham has posted instructions on installing RStudio to the stage 2 induction Canvas page.

Suggested study times have been attached to each of the three activities below. However, we do not expect everyone to complete all tasks within the 90 minute practical session – you must make every effort to complete all unfinished work by the end of this week.

1 Simulating random variables: R's built-in functions

Instructions: Try out **all** of these questions relating to material in Chapter 3 of the course notes.

You will be taken to your breakout room to complete this work, in which you can ask your tutor (Lee/Amy/Kieran/Nicola) for any help you might need.

Suggested study time: Around 45 minutes.

1. [3 Minutes] Quick Quiz. Match the following scenarios to the correct function in R.

Scenario	R function
Amy wants to randomly generate random values from a Binomial distribution.	
Kieran wants to find a quantile from the Normal distribution.	
Nicola wants to find a cumulative probability from the Poisson distribution.	
Lee wants to find the density at a particular value for the exponential distribution.	

Functions to choose from:

dnorm pexp rbinom dexp rpois rnorm ppois qnorm

2. [7 Minutes] Quick Quiz.

- (a) Use R's built-in function `rexp` to simulate 10,000 values from X , where $X \sim \text{Exp}(4)$.

Write down:

- The mean of your sample: _____
- The standard deviation of your sample: _____

- (b) Suppose $Y \sim \text{Po}(6)$.

- Use R's built-in function `dpois` to find $\Pr(Y = 5)$: _____
- Use R's built-in function `ppois` to find $\Pr(Y < 4)$: _____
- What is $\Pr(Y \leq 3)$? _____

- (c) Suppose $Z \sim \text{Geom}(0.7)$.

- Use R's built-in function `dgeom` to find $\Pr(Z = 2)$: _____
- Use R's built-in function `pgeom` to find $\Pr(Z \leq 4)$: _____

3. [15 minutes]

- (a) 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 .
- (b) 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

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

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

4. [20 minutes] Write an R function to generate samples from a 3-component normal mixture distribution with parameters

$\mu_1 = 2,$	$\sigma_1 = 1,$	$w_1 = 0.3,$
$\mu_2 = 5,$	$\sigma_2 = 2,$	$w_2 = 0.4,$
$\mu_3 = 7,$	$\sigma_3 = 0.5,$	$w_3 = 0.3.$

Simulate a sample of size 1000 from this distribution and plot a histogram. *[Huge hint: See the code on page 33 of the course notes!]*

2 Simulating Random Variables: The Inverse CDF Method

Instructions: Try out **all** of these questions relating to material in Chapter 4 of the course notes.

1. [25 minutes] The aim of this question is to illustrate how the inverse CDF method works. This question is **not** typical of what you would see in an assignment or the exam, but you should still work through all parts really carefully to help gain a full understanding of the material in Chapter 4 of the course notes.

In this example we will pretend that the built-in function in R to simulate from an exponential distribution (`rexp`) does not exist. Therefore, we must use the inverse CDF method to generate random values from the exponential distribution.

- (a) Suppose $X \sim \text{Exp}(\lambda)$. Then we know that the probability density function (PDF) is given by

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

Use the following code to produce a plot of the PDF for X when $\lambda = 2$.

```

1 #Set up a sequence of values to take the x co-ordinates in our plot:
2 x = seq(0, 4, 0.01)
3 #Now generate the corresponding y co-ordinates:
4 lambda = 2
5 y = lambda*exp(-lambda*x)
6 plot(x, y, type="l", col="purple", lwd=2, main="PDF for X, where X~
   exp(2)", ylab="Density")

```

- (b) Now we want to be able to produce a sample of values from X , where $X \sim \text{Exp}(2)$, and we do this using the inverse CDF method. Recall that we must:
 - Find the inverse of the CDF (cumulative distribution function) for our target distribution
 - Apply the inverse CDF to $U \sim U(0, 1)$

Let's see why this works. First, we need to find the CDF. You might know from other courses (e.g. MAS1604, MAS2904) that the CDF for the exponential distribution is

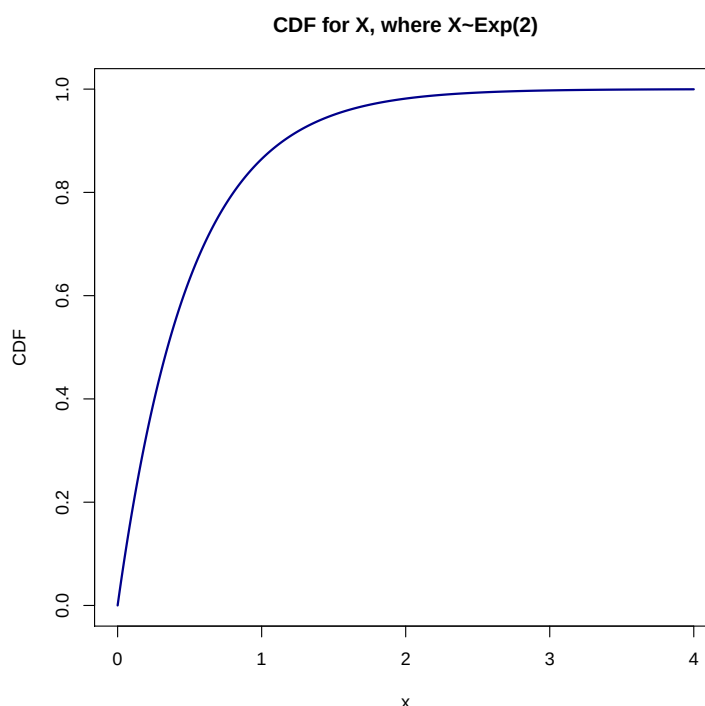
$$F_X(x) = 1 - e^{-\lambda x}, \quad x > 0, \quad \lambda > 0.$$

But what if you didn't know this? Then you'd need to integrate over the PDF (as the CDF gives areas under the curve of the PDF). Thus,

$$\begin{aligned}
 F_X(x) &= \int_0^x \lambda e^{-\lambda t} dt \\
 &= \lambda \left[-\frac{1}{\lambda} e^{-\lambda t} \right]_0^x \\
 &= [-e^{-\lambda t}]_0^x \\
 &= (-e^{-\lambda x}) - (-e^{-0}) = 1 - e^{-\lambda x},
 \end{aligned}$$

as required.

- (i) Can you explain why the integration limits are $[0, x]$?
 - (ii) Why have we introduced a 'dummy variable' t ?
- (c) Let's see if the CDF we derived in part (b) makes sense. Remember, this should give us areas under the curve in the plot you produced in part (a).
- (i) Find $F_X(x)$ for $x = \{0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0\}$, when $\lambda = 2$ (by hand using a calculator, or in R)
 - (ii) Compare your answers in part (i) to your plot in part (a) of this question. Do these answers make sense, with reference to the plot?
- (d) Follow the code used in part (a) to produce a plot of the CDF for X , where $X \sim \text{Exp}(2)$. You should get something which looks like the plot shown below.



- (e) Look carefully at the plot of the CDF above. If we want to simulate values from $X \sim \text{Exp}(2)$, then we want values lying on the x-axis. Notice also that values from the CDF, shown on the y-axis, are between 0 and 1 (as they represent (cumulative) probabilities). So now it might become apparent why the inverse CDF method works:
- If we simulate values u from $U \sim U(0, 1)$, this in effect gives us random values from the y-axis in our plot in part (d)
 - To then obtain corresponding values from X , on the x-axis, we need to feed our values from U through the *inverse* of the CDF
 - In other words, to get from x-axis to y-axis, we apply the CDF F_X (to x) to get values between 0 and 1 (u); thus, to get from y-axis to x-axis we must apply the inverse CDF F_X^{-1} (to u) to get values from the target distribution (x)

- (f) By hand, invert the CDF to show that

$$F_X^{-1}(x) = -\frac{1}{\lambda} \log(1 - x).$$

- (g) Now try out the following code, which applies the inverse CDF for the exponential distribution to simulated values u from $U \sim U(0, 1)$ – in effect getting us from y-axis to x-axis in the plot you produced in part (d).

```
1 exp.sim = function(n, lambda){
2   U = runif(n)
3   answer = -1/lambda*log(1-U)
4   return(answer)}
```

Make sure you try out the above code for `lambda = 2` and use a big value for `n`. Store your output in `sims`.

- (h) Now produce a histogram of the sample you generated in part (g), and superimpose the PDF just to check, using the following code:

```
1 hist(sims, prob=TRUE, main="Inverse CDF method for Exp(2)", ylab="
  Density")
2 x = seq(0, 4, 0.001)
3 lambda = 2
4 y = lambda*exp(-lambda*x)
5 lines(x, y, type="l", lwd=2, lty=2, col="darkred")
```

2. [20 minutes] Aim: Write an R function that uses the inverse CDF method to generate a sample from the distribution with probability density function

$$f_X(x) = \begin{cases} \left(\frac{2}{x}\right)^3 & \text{when } x > 2, \\ 0 & \text{otherwise.} \end{cases}$$

- (a) Note that you are given the PDF f_X , not the CDF F_X . Thus, you should obtain the CDF by integrating, i.e. find

$$\int_2^x \left(\frac{2}{t}\right)^3 dt.$$

- (b) Find the inverse of the CDF, that is, F_X^{-1} .
- (c) Write an R function called `inv.cdf` that takes a single input variable: `n`, the number of realisations to generate. Within this function, you should implement the inverse CDF method to generate values from X by applying your answer for F_X^{-1} in part (b) to observations generated randomly from a $U(0, 1)$ distribution.
- (d) Execute the following code to produce a histogram of 1000 realisations from X , on a density scale, and to superimpose the PDF of X :

```
1 X = inv.cdf(1000)
2 hist(X, probability = TRUE)
3 curve((2/x)^3, from = 0, to = 100, col = "red", add = TRUE)
```

3. [20 minutes] Annual maximum sea levels (in feet), observed at locations on the US Gulf coast, are assumed to follow a *Gumbel* distribution with cumulative distribution function

$$F_X(x) = \exp \left\{ - \exp \left(- \left[\frac{x - \mu}{\sigma} \right] \right) \right\}, \quad \mu \in \mathbb{R}, \quad \sigma > 0.$$

- (a) Show that the inverse of the CDF is given by

$$F_X^{-1}(x) = \mu - \sigma \log(-\log x).$$

- (b) Write an R function called `gumbel.sim` that takes three input variables: `n`, the number of realisations to generate; `mu`, the value of the Gumbel location parameter; and `sigma`, the value of the Gumbel scale parameter. Within this function, you should implement the inverse CDF method to generate values from X by applying your answer for F_X^{-1} in part (a) to observations generated randomly from a $U(0, 1)$ distribution.
- (c) From annual maximum sea levels observed at New Orleans, maximum likelihood estimates for the Gumbel parameters are obtained as $\hat{\mu} = 8.36$ and $\hat{\sigma} = 1.20$. With these estimates, use your function in part (b) to simulate 10,000 years of sea level maxima for this location. Store your realisations, and produce numerical/graphical summaries of your realisations.
- (d) Use your simulated series in part (c) to estimate the probability that sea levels at New Orleans will exceed 16 feet in any given year.

Solutions: Part 1

1. Amy – rbinom; Kieran – qnorm; Nicola – ppois; Lee – dexp
2. (a) $x = \text{rexp}(10000, 4)$: Both $\text{mean}(x)$ and $\text{sd}(x)$ should be around $1/4$
 (b) $\text{dpois}(5,6) = 0.161$; $\text{ppois}(3,6) = 0.151$; $\Pr(Y \leq 3) = \Pr(Y < 4) = 0.151$
 (c) $\text{dgeom}(1, 0.7) = 0.21$; $\text{pgeom}(3, 0.7) = 0.992$
3. (a) (i) $\text{dbinom}(3, 12, 0.01)$
 (ii) $\text{pbinom}(2, 12, 0.01)$
 (iii) $x = \text{rbinom}(1000, 12, 0.01)$; $\text{mean}(x)$ should be close to $n \times p$; $\text{var}(x)$ should be close to $n \times p \times (1 - p)$
 (b) (i) No - why have such definite cutoffs, and equal probability in between?
 (ii) $Y = \text{runif}(1000, -15, 45)$; $\text{mean}(Y)$; $\text{var}(Y)$
 (iii) $\text{hist}(Y, \text{xlab} = \text{"Minutes late"})$
4. Simple amendment/extension of the code on page 33 of the notes:

```

1 normal.mixture = function(n, mu1,sig1,w1, mu2,sig2,w2, mu3,sig3,w3) {
2   p = c(w1, w2, w3)
3   x = vector(mode = 'numeric', length = n)
4   for (i in 1:n) {
5     j = sample(c(1, 2, 3), 1, prob = p)
6     if (j==1) {
7       x[i] = rnorm(1, mu1, sig1)
8     }
9     else if (j==2){
10      x[i] = rnorm(1, mu2, sig2)
11    }
12    else {x[i] = rnorm(1, mu3, sig3)}
13  }
14  return(x)
15 }
16 test = normal.mixture(1000, 2,1,0.3, 5,2,0.4, 7,0.5,0.3)
17 hist(test)

```

Solutions: Part 2

1. (a) Use code as given.
- (b) (i) We use $[0, x]$ as we know that $x > 0$; we also want to find areas under the curve up to and including a particular point x (as the CDF gives us $\Pr(X \leq x)$); (ii) The 'dummy variable' t enables the final answer for the CDF to be expressed in terms of x .
- (c) (i) $F_X(x) = \{0, 0.632, 0.865, 0.950, 0.982, 0.993, 0.998\}$ for $x = \{0, 0.5, \dots, 3.0\}$; (ii) Yes, these probabilities seem to correspond to areas under the PDF to the left of x
- (d) `x = seq(0, 4, 0.001); lambda = 2; y = 1-exp(-lambda*x);
plot(x, y, type="l", main="CDF for X, where X Exp(2)", ylab="CDF")`
- (e) No solution needed.
- (f) Set $y = 1 - e^{-\lambda x}$ and rearrange to make x the subject, giving the answer as stated.
- (g) `sims = exp.sim(10000, 2)`
- (h) No solution needed.
2. (a) We find

$$\begin{aligned}
 F_X(x) &= \int_2^x \left(\frac{2}{t}\right)^3 dt \\
 &= 2^3 \int_2^x t^{-3} dt \\
 &= 8 \left[-\frac{1}{2t^2} \right]_2^x \\
 &= 1 - \frac{4}{x^2},
 \end{aligned}$$

giving

$$F_X(x) = \begin{cases} 0 & x \leq 2 \\ 1 - \frac{4}{x^2} & x > 2 \end{cases}$$

- (b) Finding the inverse - rearranging F_X to make x the subject gives

$$F_X^{-1}(x) = \sqrt{\frac{4}{1-x}}, \quad x > 2.$$

- (c) R function:

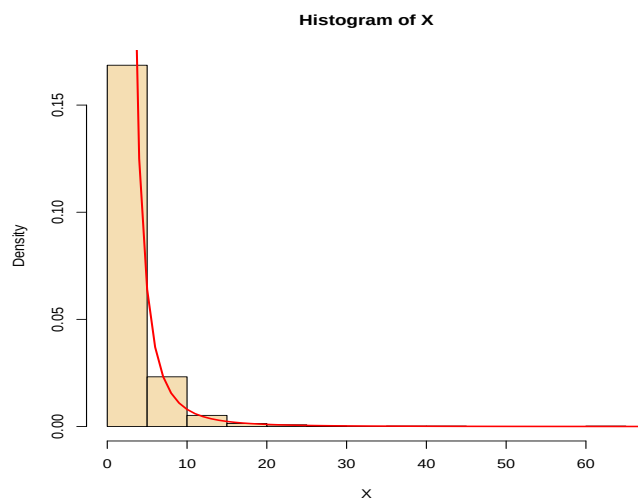
```

1 inv.cdf = function(N){
2   U = runif(N)
3   answer = sqrt(4/(1 - U))
4   return(answer)}

```

(d) Running/testing the function:

```
1 x = inv.cdf(1000)
2 hist(x, prob=TRUE, main="Histogram of X")
3 curve((2/x)^3, from=0, to=70, add=TRUE, col="red")
```



3. (a) Let $y = \exp \left\{ -\exp \left(-\left[\frac{x-\mu}{\sigma} \right] \right) \right\}$. Then

$$x = \mu - \sigma \log(-\log y),$$

and so $F_X^{-1}(x)$ is as stated.

(b) R code:

```
1 gumbel.sim = function(n, mu, sigma){
2   U = runif(n)
3   answer = mu-sigma*log(-log(U))
4   return(answer)}
```

(c) Running/testing the function:

```
1 test = gumbel.sim(10000, 8.36, 1.20)
2 mean(test); var(test)
3 hist(test)
```

(d) `length(test[test>16])/10000`