

Mini Project (Problem Solving Exercises 2)

Solutions should be submitted to the assignment submission area in Canvas by **4pm on Friday 6th December 2024**.

Your solutions to all questions should be typed and should form a single PDF document for submission. In addition to the maximum marks shown below for each question, up to 10 marks will be awarded for style and presentation to give a total maximum of **[85 marks]**.

You should make a front cover for your work, which clearly states your name and student number. If we cannot see a student's name on any work, a mark of zero will be attached to that student.

Note that, if you want, you can work in pairs for this project. Both students should submit an identical piece of work in Canvas, which **MUST** clearly state the names of both students on the front cover.

Question 1 [Total 25 marks]

In this question, you are required to use Monte Carlo integration to approximate

$$\mathcal{I} = \int_0^{\frac{a-1}{b}} \frac{b^a}{(a-1)!} x^{a-1} e^{-bx} dx,$$

for $x > 0$ and where $a, b \in \mathbb{Z}^+$.

- (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:

```
1 set.seed(LOGIN_ID)
2 a = sample(2:10, 1); b = sample(3:12, 1)
```

where LOGIN_ID is your computer ID without the 'a' or 'b' at the start. If you are working as a pair, just use the LOGIN_ID for one student in the pair. In your solutions, write down your values for a and b and the resulting integrand $f(x)$. For example, if your generated values for a and b were 7 and 3 respectively, you would write down:

$$a = 7; \quad b = 3; \quad f(x) = \frac{243}{80} x^6 e^{-3x}.$$

- (b) Use R to produce a plot of $f(x)$, experimenting with `xlim` and `ylim` to make sure your plot is nicely displayed. Also use the `abline` command to indicate a suitable *simulation grid* over which you will perform your Monte Carlo integration. Include this plot in your solutions, with appropriately labelled axes and title. **[5 marks]**
- (c) Using a `for` loop over $N = 10^6$ replications, estimate your integral using Monte Carlo Integration. 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. **[17 marks]**

- (d) Use a built-in command available in R to verify your answer to part (c). Include the R code in your solutions document. *Hint: the function over which you are integrating is actually the density function for a standard probability distribution. Can you spot which distribution this is?* **[3 marks]**

Question 2 [Total 25 marks]

Suppose that X is a random variable with probability density function (PDF)

$$f_X(x) = \frac{x}{\sigma^2} e^{-x^2/(2\sigma^2)}, \quad \text{for } x > 0,$$

where σ is a fixed constant.

- (a) Find the cumulative distribution function (CDF) for X and write a function to generate simulated values of X using the inverse CDF method. Include all necessary calculations in your solutions; also include the R code for the function that you write. **[12 marks]**
- (b) Let d be the last digit of your student ID. Let $\sigma = 1 + (d/2.5)$. Write down your value of d , and plot a density histogram of a sample of X of size 10,000 using your value of σ . Include this histogram in your solutions; again, make sure axes are appropriately labelled. If you are working as a pair, just use the ID for one student in the pair. **[5 marks]**
- (c) Check that your function has worked by superimposing a curve on your histogram showing the PDF. **[2 marks]**
- (d) The distribution you have simulated from is often used in the physical sciences to model wind speeds and wave heights; it is also used to model noise variation in magnetic resonance imaging (MRI). Can you find the name of this distribution? **[2 marks]**
- (e) Use your simulated values of X to verify the expectation and variance of the distribution you identified in part (d). **[4 marks]**

Question 3 [Total 25 marks]

If the random variable X follows a *Beta distribution*, that is, $X \sim \text{Be}(a, b)$, then it has PDF

$$f_X(x) = \frac{x^{a-1}(1-x)^{b-1}}{B(a, b)}, \quad 0 < x < 1, \quad a, b \in \mathbb{Z}^+,$$

and where $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ and $\Gamma(a) = (a-1)!$.

Suppose a is the value you generated in question 1, and $b = 1$. Suppose also that $Y = -\log X$.

(a) Using R:

- (i) Produce a random sample of size 10,000 from X using a single line of code. Include this code in your solutions. **[2 marks]**
- (ii) Write an R function that produces a random sample from Y . Include this function in your solutions. **[4 marks]**
- (iii) Use your R function in part (ii) to produce a random sample of size 10,000 from Y . In your solutions, include histograms of your random samples from X and Y side-by-side, in a 1×2 panel of plots. **[4 marks]**

(b) By hand:

- (i) Derive the PDF for Y , and include full workings in your solutions. **[6 marks]**
 - (ii) Find the exact value for $P(Y > E[Y])$. **[5 marks]**
- (c) Use your simulations in part (a)(iii) to find an approximation to $P(Y > E[Y])$, and compare to your answer to part (b)(ii). Include any relevant R code/output in your solutions. **[4 marks]**

Solutions

Question 1 (25 marks)

- (a) Students have randomly generated values for a and b linked to their student numbers. Suppose $a = 6$ and $b = 11$; then

$$\mathcal{I} = \int_0^{5/11} \frac{11^6}{5!} x^5 e^{-11x} dx. \text{ [No marks]}$$

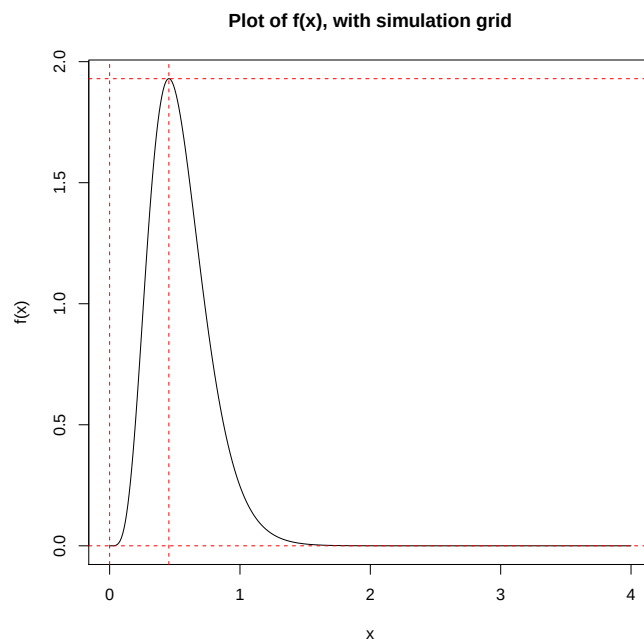
- (b) [5 marks] – no marks for code; [3] for plot; [1] for simulation grid; [1] for suitable labels.

Note that the vertical lines in the simulation grid should be very similar for everyone; the horizontal line going over the top of the plot can be anywhere provided it covers the plot (it needs to be placed at at least the maximum value of the function). Deduct marks accordingly if the range of values on the x-axis is not appropriate - for example, if students show the curve for $x < 0$ or have not extended the plot enough to see the curve return to the x-axis.

```

1 #Setting up:
2 a = sample(2:10, 1); b = sample(4:12, 1)
3 #Upper limit for integration:
4 upper = (a-1)/b
5 #Graph:
6 x1 = seq(0, 4, 0.01)
7 y1 = ((b^a)/gamma(a))*(x1^(a-1))*exp(-b*x1)
8 plot(x1, y1, type="l", main="Plot of f(x), with simulation grid", xlab=
   "x", ylab="f(x)")
9 top = max(y1)
10 abline(h = top, lty=2, col="red")
11 abline(v=0, lty=2, col="red")
12 abline(v=upper, lty=2, col="red")

```



- (c) [17 marks] – [2] for correct generation of points (lines 5-6 in the below code); [2] for evaluation of function (line 8); [4] for if statement and counter (lines 10-12); [2] for calculation of P (line 15); [2] for adjustment by area of simulation grid (line 17); [2] for correct overall structure of the main function; [1] for inclusion of comments (see green comments as examples below); [2] for correct implementation and answer (lines 20-21; check against answer to part (d))

```

1 MCI = function(N){
2   no_of_hits = 0
3   for(i in 1:N){
4     #Throw random points at the simulation grid:
5     x = runif(1, 0, upper)
6     y = runif(1, 0, top)
7     #Evaluate the function at each x-value:
8     f = ((b^a)/(factorial(a-1)))*(x^(a-1))*exp(-b*x)
9     #If the y-value falls under the curve, increase counter by 1!
10    if(y<f){
11      no_of_hits = no_of_hits+1
12    }
13  }
14  #Find the proportion of points under the curve:
15  P = no_of_hits/N
16  #Adjust by the area of the simulation grid:
17  area_under_curve = P*(upper*top)
18  return(area_under_curve)}
19
20 MCI(10^6)
21 [1] 0.383869

```

- (d) [3 marks] Note that students MUST include the R code and the answer. Zero marks if the incorrect function is used in R.

```

1 pgamma(upper, a, b)
2 [1] 0.3840393

```

Question 2 (25 marks)

- (a) **Cumulative Distribuiton Function:** [4 marks] Note that most students are likely to use integration by substitution here. Deduct marks as appropriate if insufficient explanation or working is provided.

$F_X(x) = 1 - e^{x^2/(2\sigma^2)}$ since differentiating this gives $f_X(x)$ and $F_X(\infty) = 1$.

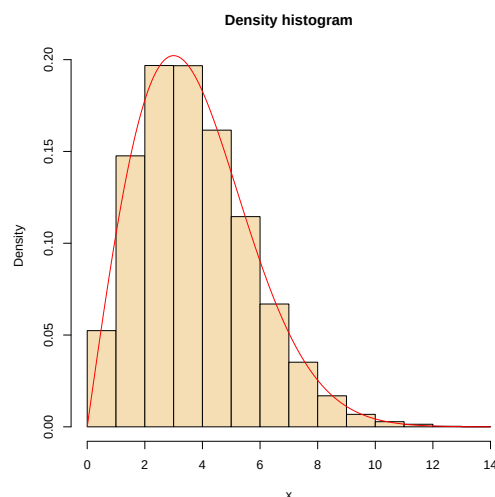
Inverse and R function: [8 marks] – [1] for generation of u (line 2 in code below); [4] for correct inverse function, with application to u (line 3); [3] for correct overall structure of function, with `return` statement etc. Deduct [2] marks if insufficient working provided for inverse. Note that students' values for σ will vary depending on their student number. Note also that line 5 below is not needed yet - this relates to part (b) of the question.

$F_X^{-1}(x) = \sqrt{-2\sigma^2 \log(1-x)}$ so the function is:

```
1 inv.cdf = function(N, sigma){
2   u = runif(N)
3   x = sqrt(-2*(sigma^2)*log(1-u))
4   return(x)}
5 test = inv.cdf(10000, 3)
```

- (b) [5 marks] – full marks for histogram only, with appropriate labels on axes. Deduct [1] mark if labels seem inappropriate, and a further [1] mark if the histogram has not been produced on the density scale. No code needed for marks, and no marks at this stage for overlaid PDF.

Suppose student ID ends with 5, so $\sigma = 3$. Then line 5 of the above code can be used to implement the function, and we can produce a histogram of our sample:



Sample mean should be approximately $1.25 \times \sigma$. So, for $d = 5$ get approximately 3.75.

- (c) [2 marks] No code needed, check that their overlaid PDF matches the histogram (see part (c)).
- (d) Rayleigh distribution [2 marks] for name of distribution
- (e) Check `mean(test)` and `var(test)` against theoretical values (e.g. as found from Wikipedia) [4 marks] check their theoretical expectation/variance match those obtained through simulation.

Question 3 (25 marks)

(a) Suppose $a = 7$.

(i) In R: [2 marks]

```
1 X = rbeta(10000, 7, 1)
```

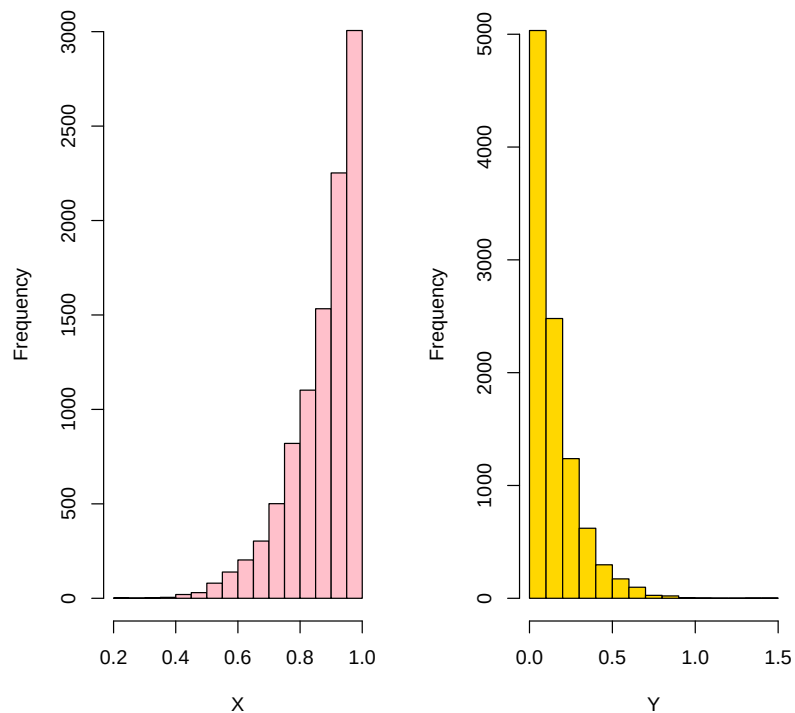
(ii) In R: [4 marks] Note that students can call the function anything they like!

```
1 transform = function(N){
2   X = rbeta(N, 7, 1)
3   Y = -log(X)
4   return(Y)}
```

(iii) In R: [4 marks] for plots Deduct 1 mark if the plots aren't side-by-side.

```
1 test = transform(10000)
2 par(mfrow=c(1,2))
3 hist(X, xlab="X", col="pink", main="Histogram of 10,000 Beta(7, 1)
   realisations")
4 hist(test, xlab="Y", col="gold", main="Histogram of 10,000
   realisations from Y")
```

Histogram of 10,000 Beta(7, 1) realisations Histogram of 10,000 realisations from Y



- (b) (i) Generally, suppose we have that $X \sim \text{Beta}(a, 1)$ (each student has their own value for $a - 7$ is used in part (a) of this question). Then

$$f_X(x) = \frac{x^{a-1}(1-x)^{1-1}}{B(a, 1)} = \frac{x^{a-1}}{\frac{1}{a}} = ax^{a-1}.$$

Now

$$F_Y(y) = \Pr(Y \leq y) = \Pr(-\log X \leq Y) = \Pr(X \geq e^{-y}) = 1 - \Pr(X \leq e^{-y}) = 1 - F_X(e^{-y}).$$

Then

$$f_Y(y) = \frac{d}{dy} (1 - F_X(e^{-y})) = -f_X(e^{-y}) \cdot -e^{-y} = a(e^{-y})^{a-1}e^{-y} = ae^{-ay},$$

and so $Y \sim \text{Exp}(a)$. **[6 marks]**. Deduct [1] mark if students do not recognise that this is an Exponential distribution (they must explicitly state this).

- (ii) Therefore, we know that $E(Y) = 1/a$ and that

$$F_Y(y) = 1 - e^{-ay},$$

and so

$$\Pr(Y > E(Y)) = 1 - \Pr(Y \leq E(Y)) = 1 - [1 - e^{-a \cdot (1/a)}] = e^{-1}.$$

[5 marks] Students might choose to integrate to get $E(Y)$, which is fine (but much more work!). They might also choose to integrate to get the CDF; also fine! They MUST state the answer precisely as e^{-1} to get full marks – deduct [2] marks if they don't do this!

- (c) In R: **[4 marks]**; [2] for R code, [1] for correct answer from R code, [1] for comment.

Students might use slightly different code to obtain this answer.

```
1 length(test[test > (1/7)]) / 10000
2 [1] 0.3699
3 exp(-1)
4 [1] 0.3678794
```

As we would expect, our simulations give an answer very close to the exact answer!

Presentation (10 marks)

Students can gain an extra 10 marks for presentation. As a guide, for students gaining full marks for presentation:

- The work is presented immaculately, with clear structure and good comments
- All code is presented extremely well, with indentations as appropriate and good use of comments where appropriate (but students should not be penalised for not including unnecessary R code)
- All plots are extremely clear – not too small, with legible (and appropriate) axis labels and titles
- Full workings are presented (where appropriate) – these can be hand-written, but must be very neat, clear and easy to read
- Students have included a cover sheet, including their name(s) and student number(s)

Other Notes

- There is no need to check student numbers and their corresponding random values generated for a , b , d and σ . All students' answers should be roughly the same, regardless of the values these parameters!
- Students were permitted to work in pairs. Some might have done this, some might not have done this, it was up to them and we didn't take a record of this. All students were told to upload a submission to Canvas, so please check the names on the front covers they have provided and where students have worked in pairs, only mark ONE of these submissions. Both students should have uploaded the same document, and both students will be awarded the same mark.