

Probability distributions in R

Getting started

Log-in to a PC in the usual way and open Rstudio. Once you've opened Rstudio you should see four panels – if there is no text editor open in the top-left panel, click on the green button with a white plus symbol and open a new R Script. Remember – to execute a line of code in the text editor, hit Ctrl-Enter.

In the questions below, the important part is **understanding** what's going on rather than just typing the R commands. Make sure you answer the questions under each piece of R code. If you can't answer the question, then ask! Where possible, write your answer down on the worksheet. You should create the folder `Practical4` in a directory called `MAS2906` – this is where you can save your R session and any other files associated with this practical.

Question 1

Quick Quiz. Match the following scenarios to the correct function in R (see a list of functions to choose from below the table).

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

Question 2

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 $\Pr(Z \leq 4)$: _____

Question 3

- (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.

Question 4

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 MAS2904 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:

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

Listing 1: 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?

Please turn over for question 5

Question 5

The survival times of patients (X years) receiving a new cancer drug are assumed to follow a *Weibull* distribution with probability density function

$$f_X(x; \lambda, \kappa) = \frac{\kappa}{\lambda} \left(\frac{x}{\lambda}\right)^{\kappa-1} e^{-(x/\lambda)^\kappa}, \quad x \geq 0, \quad \lambda, \kappa > 0,$$

where λ is referred to as the scale parameter and κ is the shape parameter.

Past studies from a similar drug indicate that $\lambda = 2$ and $\kappa = 1.5$.

- (a) Can you find out if there is a built-in function in R to generate samples from the Weibull distribution? (Hint: There is! But what is the name of this built-in function?!)
- (b) Simulate 10,000 realisations from X , and plot a histogram of these values.
- (c) Use your simulated values in part (b) to estimate
 - (i) the median survival time of patients receiving this new cancer drug;
 - (ii) the probability that a randomly chosen patient receiving this new drug will survive beyond 5 years.
- (d) Now use built-in functions associated with the Weibull distribution to verify your answers in part (c).