

The University of British Columbia
Computer Science/Data Science 405/505 Modelling and Simulation
Assignment 3

You are encouraged to discuss these problems with one or more classmates, but please make sure to submit your own work.

Exercises

1. Suppose V has cdf $F_V(x) = 1 - e^{-x^2}$ when $x > 0$, and is 0, otherwise.
 - (a) Find the quantile function for V , and write an R function called `rmyV` which takes `n` as an argument and returns a vector containing `n` random variates from the distribution of V .
 - (b) Simulate 10000 values from the distribution of V and display the values in a relative frequency histogram, with the density function curve overlaid.
2. Suppose X has pdf $f_X(x) = 3x^2$, for $x \in [0, 1]$, and 0, otherwise.
 - (a) Determine the cumulative distribution function of X .
 - (b) Find the quantile function for X , and write an R function called `rmyX` which takes `n` as an argument and returns a vector containing `n` random variates from the distribution of X .
 - (c) Simulate 10000 values from the distribution of X and construct a relative frequency histogram with the graph of the pdf overlaid.
3. Suppose p is a real number in the interval $(0, 1)$, and a random variable Y has pdf

$$g(y) = pf_V(y) + (1 - p)f_X(y)$$

where f_V and f_X are defined in questions 1 and 2.

- (a) Determine the cumulative distribution function of Y .
 - (b) Write an R function called `rmyY` which takes `n` and `p` as arguments and returns a vector containing `n` random variates from the distribution of Y . (For this purpose, you will need to also use the `rbinom()` function, and the functions created in the previous exercises.)
 - (c) Simulate 10000 values from the distribution of Y , for the case where $p = 0.4$. and construct a relative frequency histogram with the graph of the pdf overlaid.
4. Consider the pdf $h(x) = |x|e^{-x^2}$, and suppose W has pdf $f_W(x) = ph(x - a) + (1 - p)h(x - b)$ for real constants a, b and $p \in (0, 1)$. Write a function called `rmyW` that takes arguments `a`, `b`, `p` and `n` and returns a vector of `n` random variates from the distribution of W . Obtain samples of 10000 W s for the cases where $a = 1$, $b = 3$, and $p = .5$, and where $a = 1$, $b = 0.5$ and $p = .3$. Plot histograms with pdf curves overlaid.

5. Consider the following scenario. A sample of $2n$ patients with a particular disease are registered in a clinical trial for a new drug therapy. The patients have been randomly assigned to two equal groups of size n : a placebo group and a treatment group. The recovery time for each patient can be modelled with a lognormal distribution with parameters μ_i and σ_i , for $i = 1, 2$, depending on which group the patient has been assigned to. All patients are recruited to the trial at the same time and the trial ends at time T , at which point, the results would be analyzed. The recovery time for any patient who has not recovered before time T would not be known; this is an example of *censoring*.
- Write a function called `rClinicalTrial` which takes `n`, `mu` (2-vector) and `sigma` (2-vector) as arguments and returns a data frame consisting of 3 columns: a column indicating the treatment group (1 or 2), a column of recovery times (some of which will not be known and should be simply recorded as T) and a column indicating whether the recovery time was censored (1) or not (0).
 - Simulate a clinical trial which should take 2 years, involving a total of 100 patients where under the placebo the parameter values are $\mu = 0.5$ and $\sigma = 1$, and under the drug treatment, the parameter values are $\mu = 0.5$ and $\sigma = 0.1$.
 - Construct side-by-side dot plots of the two groups of simulated data, highlighting the censored observations with a different plotting character from the other observations.
6. Repeat the previous question, but this time, under the assumption that patients are recruited to the study at different times - modelled as a gamma random variable with shape and scale parameters α and β . The function `rClinicalTrial` will now need additional arguments called `alpha` and `beta` but will return the same kind of data frame, where the censoring times are now the length of time the subject was in the study at time T . Run the simulation with $\alpha = 2$ and $\beta = .2$.
7. Consider the Pareto distributions of Examples 5.28 and 5.29, and suppose X is a random variable with PDF

$$f_X(x) = \frac{(k-1)}{2(1+|x|)^k}$$

where $k \in \{2, 3, \dots\}$.

- Write a function which takes `n` and `k` as arguments and returns a vector of length `n` containing simulated values from this distribution. The simplest way to do this will be to use the `rbinom()` function and the Pareto simulator to randomly assign positive or negative signs to the variates.
- For $k = 2, 3, 4$ and $k = 5$, simulate 100 samples of size 50, calculating the averages in each case. (To do this step, you should use a `for()` loop.) Construct normal QQ-plots of the 100 averages for each value of k .
- For which values of k does the central limit theorem appear to hold? What condition of the central limit theorem is violated in the other cases?