2 Data Gathering

Exercise 2.1

1. Assume that $X_1, \ldots, X_n$ are n samples from a cdf $F(x)$. For a given value $x \in \mathbb{R}$, show that $n\hat{F}(x)$ follows a binomial distribution with parameters n and success probability $F(x)$.
2. Calculate $\mathbb{E}(\hat{F}(x))$ and var$(\hat{F}(x))$.

Exercise 2.2

1. Simulate $n > 0$ independent samples of a random variable following a normal distribution with mean 0.3 and standard-deviation 1.4.
2. Compute and plot the ecdf of these samples.
3. Compare with a figure the ecdf with the true distribution.
4. Propose an algorithm to verify (approximatively) the Dvoretzky-Kiefer-Wolfowitz inequality. Code this algorithm.
   Hints : the verification should be based on M sequences of n samples. Each sequence is used to compute an estimate of the maximum gap between the ecdf and the cdf. Finally, the M sequences allow you to estimate the probability that the maximum gap exceeds a given value ε.
5. Discuss the accuracy of the Dvoretzky-Kiefer-Wolfowitz inequality with respect to n.
6. Is the empirical mean close to the true mean ? Discuss the result with respect to n.
7. Conclude.

Exercise 2.3

Assume that $r_1, r_2, \ldots, r_n$ are n samples from a uniform distribution $(0, 1)$. Use the Inverse-transform technique to calculate n samples $x_1, x_2, \ldots, x_n$ from the Triangular distribution. The probability density function of the Triangular distribution is :

$$f_a(x) = \frac{2}{a}\left(1 - \frac{x}{a}\right),\ 0 \leq x \leq a,$$

where $a > 0$ is known, and $f_a(x) = 0$ otherwise.

1. Verify that $f_a(x)$ is a probability density function. Draw $f_a(x)$. Justify the name of this distribution.
2. Calcultate the cdf $F_a(x)$.
3. Calculate $F_a^{-1}(x)$.
4. Describe the algorithm to calculate the samples $x_1, x_2, \ldots, x_n$.

Exercise 2.4

We want to generate n i.i.d. samples $x_1, \ldots, x_n$ from a discrete random variable X following the distribution :

i	1	2	3	4
u_i	1	3	5	7
$p_i = \Pr(X = u_i)$	0.1	0.3	0.4	0.2

1. Propose an algorithm which transforms n samples $r_1, r_2, \ldots, r_n$ generated from a uniform distribution $(0, 1)$ into the n samples $x_1, \ldots, x_n$.

2. Implement the algorithm with R.
3. Verify the quality of the algorithm by plotting the histogram of $n = 1000$ samples.
4. Let Y_i be the fraction of generated samples taking on the value u_i. What is the mean and the variance of Y_i ? For $n = 1000$, is the histogram an accurate estimate of the target discrete distribution ? Is the accuracy the same for all the Y_i's ?

Exercise 2.5

Let us consider a random vector (X, Y) composed of two random variables. A random vector is said to be a normal random vector if all linear combinations of X and Y is a normal random variable, i.e., if the random variable $T = aX + bY$ is a normal random variable for all $(a, b) \in \mathbb{R}^2$, $(a, b) \neq (0, 0)$.

1. Show that a couple of two independent normal random variables (X, Y), with $\mathbb{E}(X) = m_X$, $\mathbb{E}(Y) = m_Y$, $\text{var}(X) = \sigma_X^2$ and $\text{var}(Y) = \sigma_Y^2$, is a normal random vector.
2. Let U, V and W three normal random variables with zeros means and unit variances. These variables are used to create the random variables X and Y :

$$X = \sqrt{1-\varrho}\,U + \sqrt{\varrho}\,W \quad (1)$$
$$Y = \sqrt{1-\varrho}\,V + \sqrt{\varrho}\,W \quad (2)$$

   where $0 \leq \varrho \leq 1$ is a given parameter.

   (a) Show that (X, Y) is a normal random vector.

   (b) Calculate $\mathbb{E}(X)$, $\mathbb{E}(Y)$, $\text{var}(X)$, $\text{var}(Y)$, $\text{cov}(X, Y)$ and $\text{corr}(X, Y)$.

   (c) Propose an algorithm to generate a couple of correlated normal random variables with zero means, unit variances and a correlation $\varrho \geq 0$ from a couple of independent normal random variables.

   (d) How to modify this algorithm if $\varrho < 0$?

Hints in R

1. Useful R commands :

R-function	Package
ecdf	stats

2. To install and use a package (example with "stats") :

```
install.packages("stats")
library(stats)
```