Lab 1: Basic R, data preparation and visualization

In this lab, we will practice some basic usage of R and use R to visualize, and calculate some basic summary statistics related to a hand-written digit recognition data set.

Q1. Generate a vector of length 100 with normal distribution with mean 0 and variance 4.

a) Compute the sample mean, sample variance, give a histogram and evaluate the empirical probability of greater than 1.
b) Create a 10 by 10 random matrix using the 100 random variables generated from part (a). Using apply() function to obtain the column means of the matrix.
c) Create a new vector that contains all positive numbers from the vector generated in part (a).

Q2. Using loop and create a function.

a) Using loop in R to evaluate (approximate) the expectation of negative binomial distributed random variables having the following probability mass function.

$$P(X=k)=\binom{k-1}{r-1}p^r(1-p)^{k-r} \text{ for } k=r,r+1,r+2,\cdots$$

In your approximation, consider $r=10,20$ and 30 with $p=0.2$ and 0.5 respectively.

b) Based on the R code in part (a), create an R function that can return the approximated expectation of a negative binomial distributed random variable with any given $r>1$ and $p\in(0,1)$.

Q3. Use R to visualize the probability mass functions given in question Q3, where x-axis is the realized values of the random variables ($r,r+1,r+2,\cdots$) and y-axis is the corresponding probability mass function values.

Q4. Use *image* function in R to give a heat map of the following matrix **A** with colors generated from the gray function with 32 levels. From the smallest value in **A** to the largest value in **A**, the color should changes from white to black.

$$\boldsymbol{A} = \begin{pmatrix} 1 & -1 & 1 \\ 1 & -1 & 1 \\ -1 & 1 & -1 \end{pmatrix}$$

Q5. In the remaining part of this lab, we use the **Hand-written digit recognition** data set as an example. The database was collected and analyzed in Le Cun et al. (1990). There is a total of 9298 digitized numbers in the database, where 7291 of them are used as a training set and 2549 of them are used as a test set. The images were scanned from the U.S. Mail envelopes passing through the Buffalo, N.Y. post office. The scanned images were size normalized and de-slanted to fit a 16×16 pixel box. The resulting images are in the form of 16×16 grayscale intensities of images. The training data set containing 7291 rows and 258 columns, where the first column is the digit ID (0-9), the second to the 257th column contains the 256 grayscale values and the last column is redundant. The digit ID the actual number for the image given in the corresponding row. Please also pay attention to two points below:

- Pixels are arranged row-wisely in each row of the data set;
- From the smallest values to the largest values in the matrix, the color changes from white to black.

a) Read the training data set and test data set into R as matrices. Find out the dimension of each matrix.
b) Create a frequency table to count the numbers of images corresponding to the digit ID (0-9) in training data set.
c) Create a matrix to include all the data rows that corresponding to the digit ID 0. Do the same thing for the digit ID 1 to 9.
d) Visualize the first 5 images of each digit ID using the matrices created in part (c).