AN6100
R Analytics

FINAL- EXERCISES
R Analytics

Chin Chee Kai
cheekai@ntu.edu.sg
Nanyang Business School
Nanyang Technological University

Questions

Q1: Using seed value 3993, set randValues to a vector of 1,000 random normally distributed rounded-off integers with mean 500 and standard deviation (sd) 75.

Randomly sample 15 numbers from randValues with replacement. Call this vector sampleA. Calculate and print the mean and standard deviation of sampleA.

Repeat the sampling above again with 30 numbers (sampleB) and 60 numbers (sampleC) respectively. Which sample gives the smallest sampled standard deviation?

Q2: First set seed value of 6226 (only once). Write a function bigSmall() which first generates a random vector of 5,000 uniformly distributed rounded-off integers from 20 to 80. Call this vector randB. Convert all values of randB which are larger than 65 (inclusive) into "BIG" and otherwise "small" and return this vector of strings. How many "BIG" and "small" were generated? If an integer is randomly drawn from the 5,000 integers, empirically estimate the probability of getting a "small" integer?

Questions

Q3: Using data file "AN6100-Data-3B.csv", a CSV file is to be produced to facilitate communication with high income group earning $80,000 and above, and weighing 80kg or more. The CSV should be named "highincome.csv" and must have the following columns: Hello, Income, Job Title and Country, where Hello is constructed as "Mr" or "Ms" appended with Lastname depending on whether Gender is Male or Female.

How many different countries are found?
How many male and female are in "highincome.csv"?

Q4: Using data file "AN6100-Data-3B.csv", perform divisive clustering on the data set. Cut the dendrogram to make 3 clusters.

(a) Which cluster number is the smallest?

(b) How many people are in the largest cluster? How many in the smallest cluster?

(c) Calculate the Body-Mass-Index (BMI = $\frac{Weight}{Height^2}$) of each person and save BMI in the data frame. What is the mean BMI of all?

(d) By plotting Income vs BMI and using clustering results and considering that BMI < 18.5 is unhealthy, write ONE LINE that best describes this smallest cluster.