UCS410: PROBABILITY AND STATISTICS

Laboratory Assignment – 1

You have six courses in this semester (one of these is UCS410) and there are 750 students in your class. Generate the marks of these 750 students for the six courses using the Linear Congruential Method (LCM), which is a type of random number generator. You should note that the marks will be uniformly distributed over [0, 100].

Find the mean, median and standard deviation of the all the six subjects. Also, find the mean, median and standard deviation of the sum of marks in the six courses. You have also to draw six histograms for the marks in the six courses and one for the total marks.

Study the concept of outliers in the data. Find the outliers in the six courses that you are studying based upon IQR concept (Inter quartile range).

Note1: IQR is the range between the first and the third quartiles namely Q1 and Q3: IQR = Q3 –Q1. The data points which fall below Q1 – 1.5 IQR or above Q3 + 1.5 IQR are outliers.

Laboratory Assignment – 2

Now, you have learnt the concept of generating data using random number generation, and finding some of the descriptive measures associated with the data. Based on your interest, search a suitable data from web. Understand the data completely, find the descriptive statistics associated with this data and draw the histograms for all the variables involved. You should also find the outliers existing in the data for different variables.

UCS410: PROBABILITY AND STATISTICS

Laboratory Assignment – 3

Implement assignment 1 i.e. (random generation of student data for 6 subjects and 750 students) in R-Studio. But for the generation of random numbers explore various in-built function in R (like: rnorm, runif, sample, randu, etc.) in order to inculcate more randomness in data. Write a program in R to evaluate given data on following measure and try to understand their significance:

- Measures of central tendency: (Mean, Median, Mode)
- Measures of dispersion: (Range, Mean deviation, Standard deviation, Variance, Root mean square deviation)
- Draw Gaussian plots for each subject and try to identify type of skewness and kurtosis.

(Also try to perform above stated task using rattle API in R-studio)

Laboratory Assignment – 4

Q1) There are n people gathered in a room. What is the probability that at least 2 of them will have the same birthday?

- Use an R simulation to estimate this for various n.
- Find the smallest value of n for which the probability of a match is greater than .5.
- Explore how the number of trials in the simulation affects the variability of our estimates.

Q2) A friend has a coin with probability .6 of heads. She proposes the following gambling game. You will toss it 10 times and count the number of heads. The amount you win or lose on k heads is given by (k^2 -7k)

- Plot the payoff function.
- Make an exact computation using R to decide if this is a good bet.
- Run a simulation and see that it approximates your computation in part

UCS410: PROBABILITY AND STATISTICS

Laboratory Assignment – 5 and 6

Pre-requisite: Understand the working of binomial distribution and rle() function of R.

Based upon the acquired learning try to simulate the below mentioned question.

Q1) In a selection of a sample of size 250 one by one where both defective and non-defective items are equally likely. Now perform the simulation to calculate the estimated probability of getting the same type of item 16 times in a row.

- Use an R simulation to estimate this for various values of experiment count.

Q2) In sample of size eight of question 1, estimate the probability of selecting a different type of item in each selection, that is, that will never obtain get two defective items or two non-defective items in a row.

- Also compare the estimated probability result with actual value of probability

Q3) Six animals with some names are lined up together. Calculate the probability of lineup in an order of alphabetic series with a assumption that none is having the same name.

- Also compare the estimated probability result with actual value of probability

Q4) In Question 3, let suppose 3 animals are dogs and remaining are horses. Now calculate the probability all dogs come first.

- Also compare the estimated probability result with actual value of probability

UCS410: PROBABILITY AND STATISTICS

Laboratory Assignment – 7

Q1) Simulate normal distribution values. Imagine a population in which the average height is 1.70 m with a standard deviation of 0.1. Use rnorm to simulate the height of 1000 people and save it in an object called heights.

a) Plot the density of the simulated values.
b) Generate 10000 values with the same parameters and plot the respective density function on top of the previous plot in red to differentiate it.

This plot will show you how much a sample with 10000 simulations approximate to the real normal distribution.
c) Find the 90% interval of a population with mean = 1.70 and standard deviation = .1 between 0.05 and 0.95.
d) Calculate the qvalue corresponding to every percentile in standard normal distribution.
e) Calculte the pvalues corresponding to z values ranging from 0 to 1 at an interval of 0.05.

Q2) Download the Auto.csv data set from LMS. Based on it program the following problems in R.

a) Calculate simple (linear) correlation between car price and its fuel economy (measured in miles per gallon, or mpg)
b) Create a correlation matrix by selecting each pair of columns from dataset one by one and calculate correlation between selected pairs. Fill the values in matrix named as correlation matrix.
c) Create a new dataframe, auto_num, that contains only columns with numeric values from the auto dataframe. You can do this using the Filter function. Use the cor function to create a matrix of correlation coefficients for variables in the auto_num dataframe.

d) Use the corrgram function from the corrgram package to create a default correlogram to visualize correlations between variables in the auto dataframe.
e) Create a new dataframe, auto_subset, by subsetting the auto dataframe to include only the Price, MPG, Hroom, and Rseat variables. Use the new dataframe to create a correlogram that (1) shows correlation coefficients on the lower panel, and (2) shows scatter plots (points) on the upper panel.
f) Analyze the correlation values to understand the association between pair of column datasets.

Laboratory Assignment – 8

Q1 Implement the linear regression on a regression dataset to be downloaded from LMS using the concept of training and testing in order to understand the accuracy of results using the following metric.

a) Correlation between predicted and actual value on testing part of data.
b) Accuracy metric
c) Visualization of best fit line.

Q2 Execution of the following 3 R commands will give us the data {(x(i), y(i), z(i), i = 1, 2, …,100}.

```
x<-rpois(100, 50)
y<-rpois(100, 100)
z<-rpois(100, 150)
```

Using this data:

a) Fit the linear regression model of the form $z = a + b.x + c.y$ using,
b) Fit the 3 models of the form $y = a + b.x$, $y = a + b.x + c.x^2$, and $y = a.b^x$ to this data using
c) Find the coefficient of determination, with the help of formula, for the three models and decide for the best model.