**AC52012 Research Methods – Week 3 Lab**

**Probability; Probability Distributions**

In order to complete data set, you will need to understand the commands from the RTutorial file or to read the ‘An Introduction to R’ document that can be found at <https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf>. You also need to refer to the lecture notes for understanding of probability distributions and, specifically, binomial distribution for the discrete cases and uniform and normal distribution for continuous cases.

Most of the examples in this lab are using one of the four functions:

* r*dist*() for getting random numbers with a given distribution
* d*dist*() for probability density of a given distribution
* p*dist*()for cumulative density (i.e. area under the density curve)
* q*dist*()for values of quantiles (actually, percentiles) for the given distribution

Here *dist* can be norm (for normal distribution), binom (for binomial distribution) and unif for uniform. So, for example, the above functions for a normal distribution are rnorm, dnorm, pnorm, qnorm .

1. Plot the histograms or density curves for the following random variables  
   1. A continuous random variable measuring the height of women, which has a normal distribution with the mean of 163cm and the standard deviation of 9cm.
   2. A continuous random variable measuring the bone mineral density, which has a standard normal distribution.
   3. A continuous random variable measuring the waiting times on the bus station, assuming the buses run every 15 minutes. The random variable has a uniform distribution.
   4. A discrete random variable denoting the number of heads in 20 tosses of a fair coin.

**Hint:** The easiest way of doing this is to create a vector of a large number of values for the random variable (using, for example, seq function), then use the appropriate d*dist*function to get the densities (probabilities) for this vector of values and finally to plot this using the plot function. For the cases where you need to do a histogram, use the hist or barplot function instead (it is probably easier to use barplot, see the help files).

1. Using the bone mineral density random variable, find the following:  
   1. The probability that a person has a normal bone density (above -1.00).
   2. The probability that the person has osteopenia (bone density between -2.50 and -1).
   3. The probability that a person has osteoporosis (bone density below -2.50).
   4. The 95th percentile of the bone density, which is the value of bone density such that there is 0.95 probability that a person will have such bone density or lower.
   5. The bone density such that 80% of people have that bone density or higher. In other words, find such z that there is 0.8 probability of the bone density being z or higher.

**Hint:** Use pnorm() and qnorm() functions for this.

1. Suppose we are conducting an experiment of rolling a 6-side dice 6 times and suppose a discrete random variable x represents the number of 3s that we get.   
   1. Write an R function that will, for a given roll of 6 dices (represented as a vector) return the number of 3s in it.
   2. Generate an array of all the possible different rolls of 6 dices.  
        
      **Hint:** The easiest way to do this is to use the one of the functions for getting variations, combinations and permutations of a set. One such function is permutations from the gtools package. You will first need to install this package, using install.package(‘gtools’). If R complains about not being able to find it, you might need to select the Update Packages from the Package menu. Before that, you might need to select the Update CRAN Mirror option and select one of the mirrors for packages.  
        
      **Hint 2:** The above function creates a matrix of permutations, where rows correspond to different permutations and the columns correspond to the elements of the individual permutations. To access a single permutation (say, 5th permutation in the matrix), use y[5,] if y is your array of permutations (or variations or combinations).
   3. Apply the function you wrote in part a to all the different rolls in the part b, storing the result in a vector. This will be a vector of all the values the random variable x takes for all the simple events that can be outcomes of our experiment.
   4. Find the mean, variance and standard deviation of the random variable x using the vector you obtained in c.
   5. Plot the probability histogram for the random variable x.
   6. What probability distribution does the random variable x has? How could you have plotted in R the graph in e without going through all of the above steps?
2. (**Extra Work, if you want to be ahead**) The file EPL2021.csv contains information about all of the matches played in season 2021-22 in the English Premier League. It contains various pieces of data, such as number of goals, corners, shots (total and on targets), fouls and yellow and red cards. Of all this information, we are interested in the total number of corners in each game. Home and away corners are stored in HC and AC columns.  
   1. Load this data into R. You will first need to download the above file and then to read it in using the read.csv function.
   2. Extract the total number of corners for each game. This will be the population we will be dealing with.   
        
      **Hint:** You can extract the HC and AC columns as vectors and then do vector addition to get the total number of corners.
   3. Generate an array of all possible samples of the size 5 of the above population.  
        
      **Hint:** You need to do something similar to what you did in the problem 3b above.
   4. Calculate the mean number of corners in a sample for each of the samples above.