**Module 1 Project 1**

****

**By**

**Sunil Raj Thota**

**Date: 03/08/2021**

**Title: Project 1 Report**

**ALY 6050 – Introduction to Enterprise Analytics**

**Prof. Tamir Hegazy**

**Introduction**

In this assignment, I am using R and R Studio to generate different random numbers in and execute Chi-Square Goodness of Fit test. It helps us to identify and verify how well a statistical model fits a set of observations and whether these values have its place to the probability distribution or not? This test gives an idea between the Expected values and Observed values of each problem. The random numbers are generated by using rand() function in R, be in unform distribution or non-uniform distribution or normal distribution. I will also tend to determine the type of the distribution for each problem and showcase the plots, results, and findings. In this report, I have implemented 4 Problems with a Summary Question.

**Analysis**

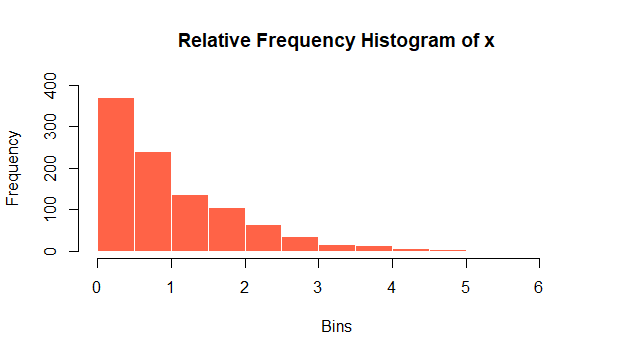
Since there is no dataset to analyze and work on Exploratory Data Analysis, we can skip this step and dive deep into the problems given and focus on them to perform some analysis. Let’s clearly understand the project requirements and go through each problem in a step-by-step manner.

**Problem 1:**

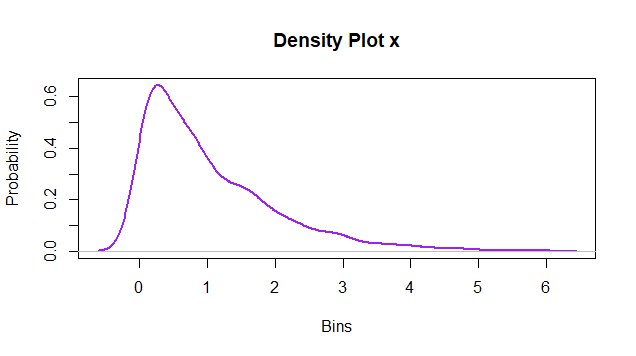
In this problem, I have to work on generating 1000 random numbers in R by using the runif() function. After this, we tale natural logarithmic values of the randomly generated values to ensure all are positive and finite.

After this, let’s plot the distribution on a histogram, density plot to see the plot characteristics. Then we will test the Chi-square Goodness of Fit test whether the values which are generated tend to a probability distribution or not? So, once the random values have been generated, we will easily able to find and calculate the x value with an exponential value of 1.

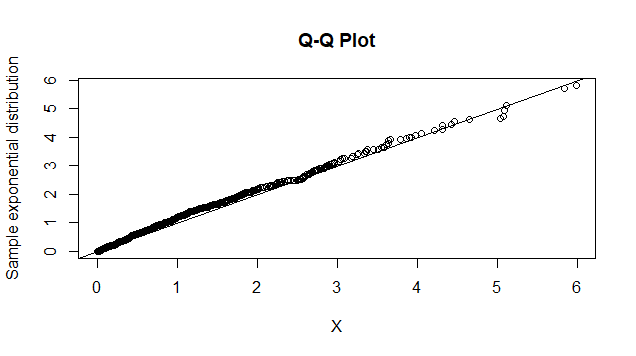
Now, let’s create a Relative Frequency Histogram plot of x and analyze the data. We can see that it is similar to the exponential distribution.



From the above plot we can say that the probability declines at beginning and slickly at the end. So, the smaller the bin, the bigger the probability will be. I also plotted the density plot to get an outline on the random values distribution. This distribution resembles like a Gamma Distribution.

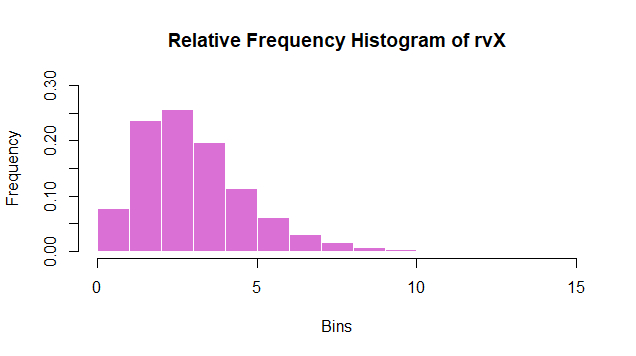


Next, I have plotted a QQ and Normal Plots



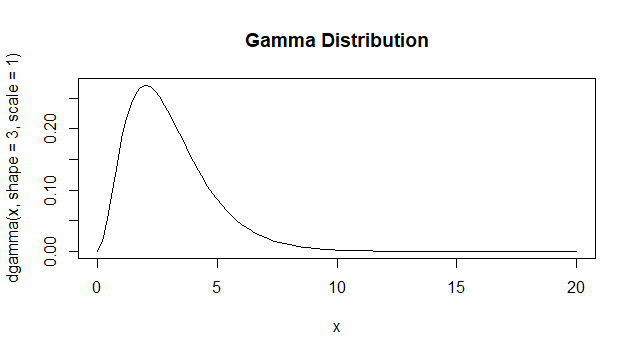
Since the QQ plot fall on the reference line and did not get deviated from it, we can conclude that x is in a normal distribution. Now, let’s perform Chi-square test on x and I found that p-value is 0.9795 which is greater than the 0.05 and it seems to be a good fit. So, I fail to reject the Ho. It is evident from the experiment that if a random sample from uniform distribution is converted taking the negative natural logarithm of the sample, then the resulting sample follows exponential distribution with mean 1.

**Problem 2:**

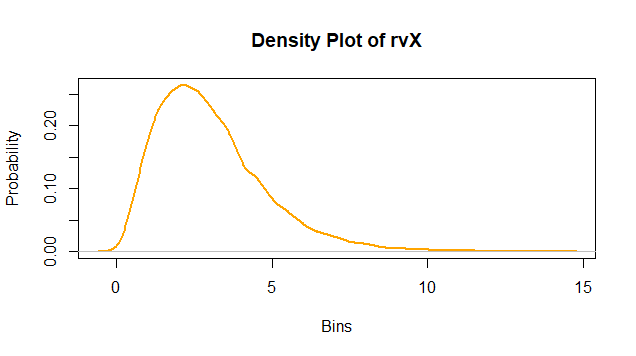


In this problem, we have to generate three different random number values of each 10000 numbers. Then let’s measure the logarithmic value by multiplying them.

From the above plot we can see the histogram of x. Likewise, let’s plot density plot and it looks like a Gamma Distribution. We can see that the curve is right-skewed.

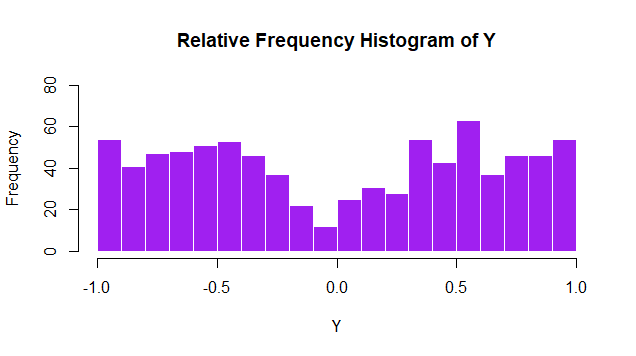


Now, let’s show probability distribution using QQ Normal plot



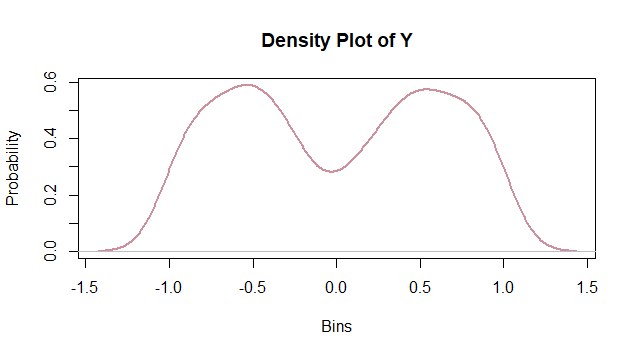
By this we can state that it is not in a normal distribution manner. We can state that from the Relative Frequency Histogram Plot, it resembles like a wave. The sum should be a Gamma random variable. The Degrees of freedom is n - 1. From the Chis-square Goodness of Fit Test, we have found out the p-Value as 0.2397, which is greater than the 0.05. So, we fail to reject the Null Hypothesis Ho. We can also state that the Gamma Distribution for n random variables will have a probability distribution. So, it perfectly fits the probability density function of Gamma Distribution.

**Problem 3:**



In this problem, I have to work on generating 1000 random numbers in R by using the runif() function for two random numbers. After this, we generate natural logarithmic values of the randomly generated values to ensure all are positive and finite. In this, we are analyzing individual vales of each random values with the logarithmic function. We also plotted a Density plot to showcase the probability distribution.

To satisfy the given conditions, In this problem I am using for loop and conditional statements to meet the logic to generate 1000 times. Then, it will generate Y values based on the given conditions. We will use these values to plot a Histogram and find the Chi-Square Goodness of Fit Test.



The p-value of the test is far less than the 0.05. So, we can state that the observed values are way different from the expected ones. Also, the sum of n exponential random variables is a Gamma. So, it is a gamma distribution. Exponential Frequency is measured by the exponential distribution function and the expected frequency values cross produced by the sample size.

**Problem 4:**

In this, as done in Problem 3, here also we will do the same thing and check for the given condition i.e., x2 < k by working on the iterations and generating the W values. After the thorough analysis we can say that it is a normal distribution and concludes with a Chi square test.

**Table for M and W**

|  |  |
| --- | --- |
| **M** | **W** |
| 10 | 0.01180638 |
| 20 | 0.02361275 |
| 30 | 0.03541913 |
| 40 | 0.0472255 |
| 50 | 0.05903188 |
| 60 | 0. 07083825 |
| 70 | 0.08264463 |
| 80 | 0.094451 |
| 90 | 0.1062574 |
| 100 | 0.1180638 |
| 200 | 0.2361275 |
| 300 | 0.3541913 |
| 400 | 0.472255 |
| 500 | 0.5903188 |
| 600 | 0.7083825 |
| 700 | 0.8264463 |
| 800 | 0.94451 |
| 900 | 1.062574 |
| 1000 | 1.180638 |

**Summary:**

1. If 𝒓 is a standard uniform random variable, then −𝑳N(𝒓) has the **Exponential** probability distribution.
2. The sum of three independent and identically distributed **3 Uniform** random variables have the **Gamma** probability distribution.
3. The output of the algorithm of problem 3 has a **Normal** probability distribution.
4. In step 2 of the algorithm of problem 3, random variables 𝑿𝟏 and 𝑿𝟐, each of whose probability distribution is **Exponential** are used to generate a random value 𝒀 that has the **Normal** probability distribution.
5. The random value 𝑾 that was discussed in problem 4, has the **Normal** probability distribution. The expected value of 𝑾 is: **1.19**

**Conclusion**

From this project I have gained a lot of knowledge and got familiar with the histograms, relative frequencies, Chi-Square Goodness Fit Test and determined the relationships and correlations between them. I previously did not work on these distributions and tests, but from this Assignment I had a great time in learning them. In coming days and assignments, I will make sure to use this knowledge and apply on the problems.

**References**

[1] Chi-squared Distribution was retrieved from <http://www.r-tutor.com/elementary-statistics/probability-distributions/chi-squared-distribution>

[2] Antoine Soetewey (2020-01-27) Chi-square test of independence in R was retrieved from <https://statsandr.com/blog/chi-square-test-of-independence-in-r/>

[3] Pearson's Chi-Squared Test For Count Data was retrieved from the <https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/chisq.test>