**Introduction to Enterprise Analytics**

# ALY6050 Module 1 Assignment

# Faculty: Prof. Richard He

# 2020 Winter CPS Quarter

# Image result for neu cps

Shraddha Gopalakrishnan [NUID: 001376444]

**Introduction**

The module one project showcases how descriptive statistics can be applied on numbers using different types of probability distributions. It focuses mainly on the random number generation process which is done to gain insights into the data in all the four sub-parts of the assignment. In each question we are asked to carry out the Chi-squared test for verifying if the randomly generated values indicate any probability distribution. There are also graphical representations which highlight the tedious calculations of the Excel. Let us jump onto the analysis of the data values and understand how important chi-squared test is in the world of statistics.

**Analysis**

**Problem 1**

Using the **=RAND ()** function of Excel, R1 is found and 1000 values are generated for this column. X which is given by the formula **-LN(R1)** is the natural logarithm value of R1. Similarly, corresponding X values for rest 999 R1 are calculated. These 1000 values which are combined as X is used to figure out the descriptive statistics from the Data Analysis Tool Pak of Excel.

The probability distribution that is best fit for X is *exponential probability distribution* as it is a widely used distribution wherein events occur continuously and independently.

For creating a histogram, we select the column X values and get a normal histogram using data analysis by putting the input and output range positions. Excel automatically creates bins and gives us the frequency or number of occurrences of each value. Relative frequency is then found by dividing each frequency value by the sum of frequencies (i.e., 1000 here).

We now need the left and right ends for the distributions which follows these formulae

**Bin1-(Bin2-Bin1)/2** & **Bin1+(Bin2-Bin1)/2** respectively. **EXPON.DIST** gives us the expected probability values which is multiplied to the sum of frequencies (1000) to provide us with the expected frequency values. Lastly, Chi-square cardinals are estimated with the usual

**(Observed-Expected) ^2 / Expected** equation.

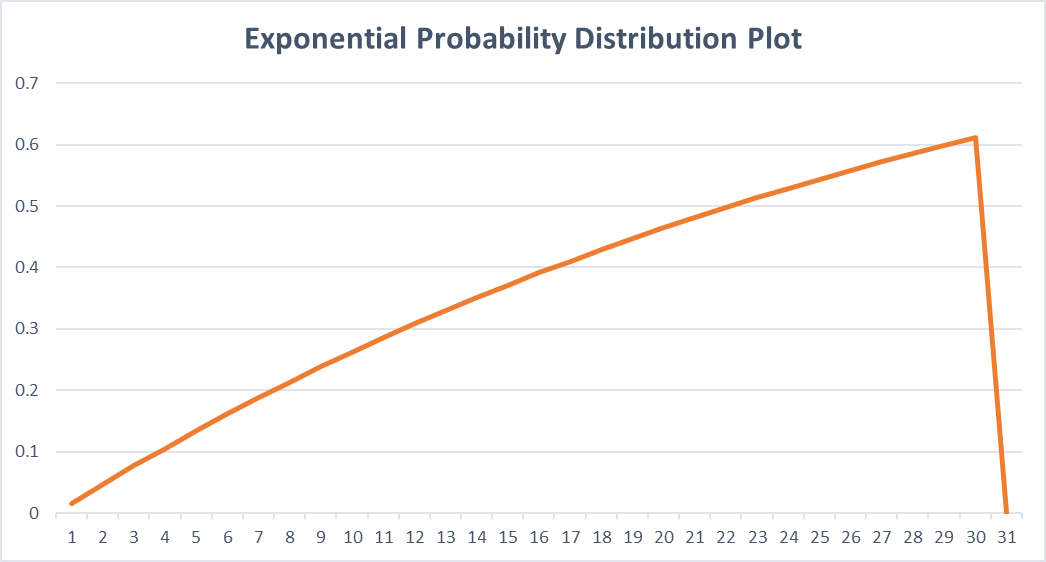
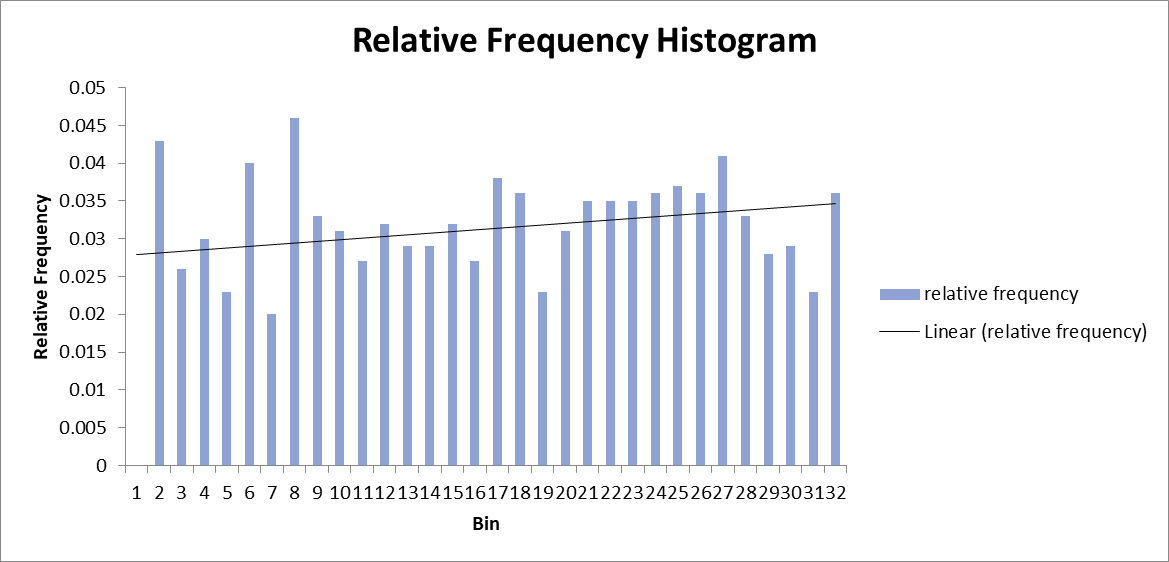
Chi-Square Goodness of the fit test calculations:

The test statistic is found by summing up all the chi-square values. Level of significance is given as 0.05 in the question. Degree of freedom is 30 (number of values) minus 2 (for last two rows left and right tail value can’t be calculated) hence our df=28. P-value is obtained using **1-CHISQ.DIST(t-stat, df, 1)** -> 1 because our frequencies are cumulative.

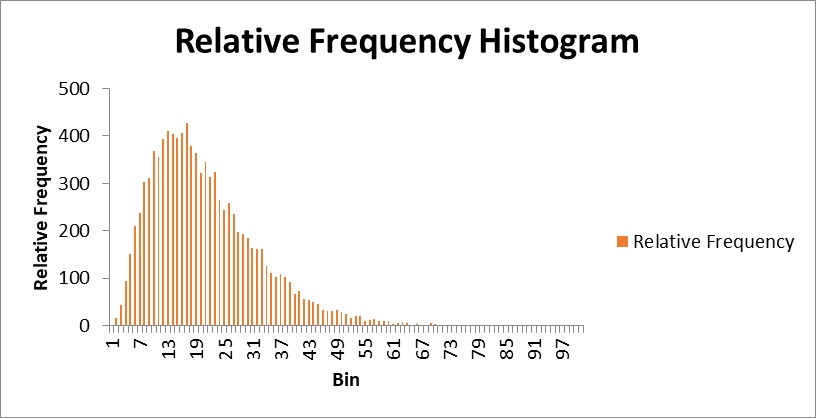
Null Hypothesis: Data is distributed exponentially

Alternative Hypothesis: Data is not distributed exponentially

Since our P-value is much lesser than the significance level, we reject the **Ho**. Therefore, our data is not exponential (can be validated also by looking at the graph).

***Graphs for 1:***

**Problem 2**

******Here for the initial step we generated continuous variables R1, R2 and R3 each comprising of 10000 values. X for all the 3 R’s is found as **X=-LN(R1R2R3)**. Similar process is followed as that in problem 1. Descriptive statistics, bins and frequency are gathered using Data Analysis Tool Pak of Excel.

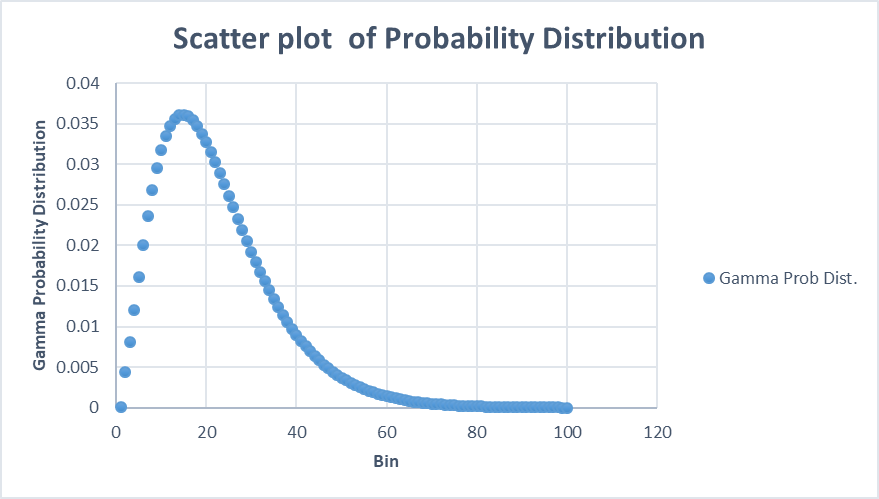
Further, a relative frequency histogram is designed using bins and relative frequency column values as input range.

The probability distribution that is best fit for X is *gamma probability distribution* as it involves waiting time events i.e., values R1, R2 and R3 separated by an average rate.

For this distribution, gamma probability density function (pdf) is hatched using **GAMMA.DIST()** function. Multiplying pdf values with 10000 easily gives the gamma frequencies. Finally, chi-square values are acquired.

Chi-Square Goodness of the fit test calculations:

The test statistic is found by summing up all the chi-square values (253.2648). Level of significance is given as 0.05 in the question. Degree of freedom is 100 (number of values) minus 2 (for last two rows left and right tail value can’t be calculated) hence our df=98. P-value is obtained using **1-CHISQ.DIST(t-stat, df, 1)**.

Null Hypothesis: Data shows gamma distribution

Alternative Hypothesis: Data does not show gamma distribution

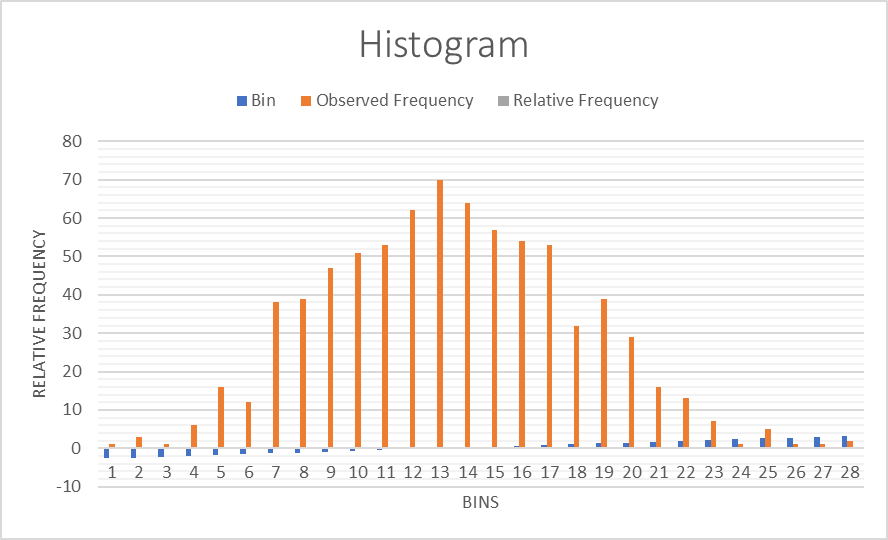
Since our P-value is less than α and so we reject the **Ho**. Therefore, our data does not show gamma distribution (Graph shows curve of gamma but that curve seems unfinished in the initial portion, hence not a gamma curve).

**Problem 3**

The 3rd part starts with generation of random variables R1-R2 with 1000 sets and their corresponding natural logarithmic value X1-X2. k is just a constant used to find an intermediate value to compare with X2. It is inferred using **(X1-1) ^2/2**. The output of the algorithm is denoted by Y whose first iteration, Y values are found using conditions as specified. If X2>=k, then a random number is generated for Y else “False” is written.

For the second iteration, if previous Y value is > 0.5, accept X1 as Y else accept -X1 as Y.

Thirdly, if X2<k, result is not attained, and the algorithm moves to the next pair of X1 and X2.

Then, descriptive statistics, Bins, Left-Right tails, Frequency and Relative Frequency are measured.

The probability distribution that is best fit for X is *normal probability distribution* as it describes how the values of the variables are distributed.

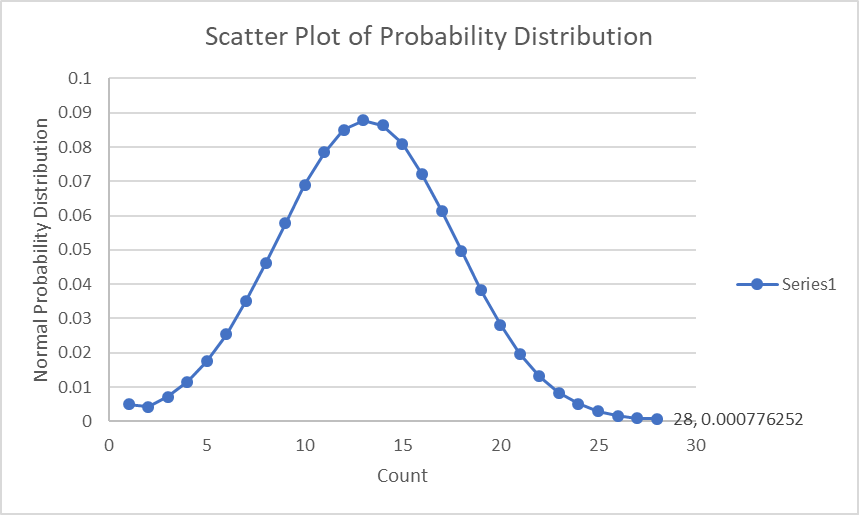
Normal probability is found using **NORM.DIST(Right-tail,Stdev,1)**. The expected frequency can now be evaluated by multiplying normal probability value with the count of accepted Y values i.e., 767.

In the end, Chi-square values are estimated using the usual equation



Chi-Square Goodness of the fit test calculations:

The test statistic is found by summing up all the chi-square values. Level of significance is given as 0.05 in the question. Degree of freedom is 28 (number of values) minus 1, so df=27. P-value is obtained using **1-CHISQ.DIST(t-stat, df, 1)** which corresponds to 1. 1 is greater than 0.05.



Null Hypothesis: Data is normally distributed

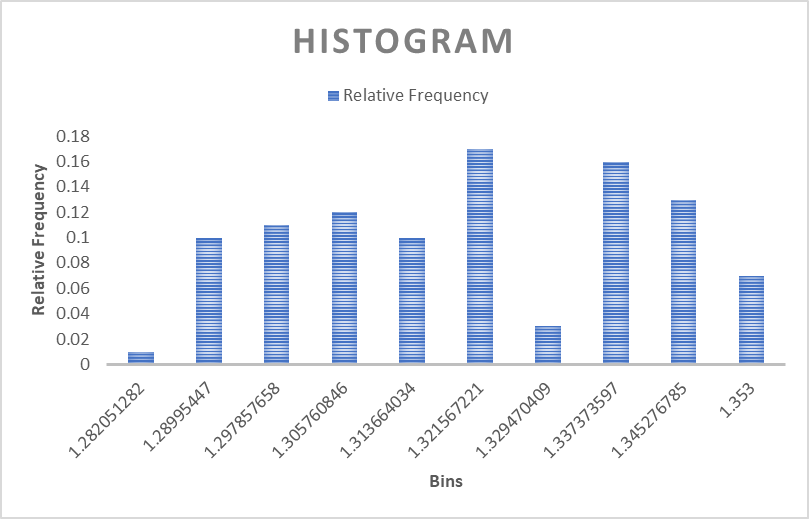
Alternative Hypothesis: Data is not normally distributed

Since our P-value is greater than 0.05, we fail to reject **Ho** and our data is normally distributed as clearly seen from the graph as well.

**Problem 4**

In this part, the Y values which were rejected in the 3rd part as “False” and “#NA” are manipulated and put to use so that the final output of Y can be useful. The changes are done as follows:

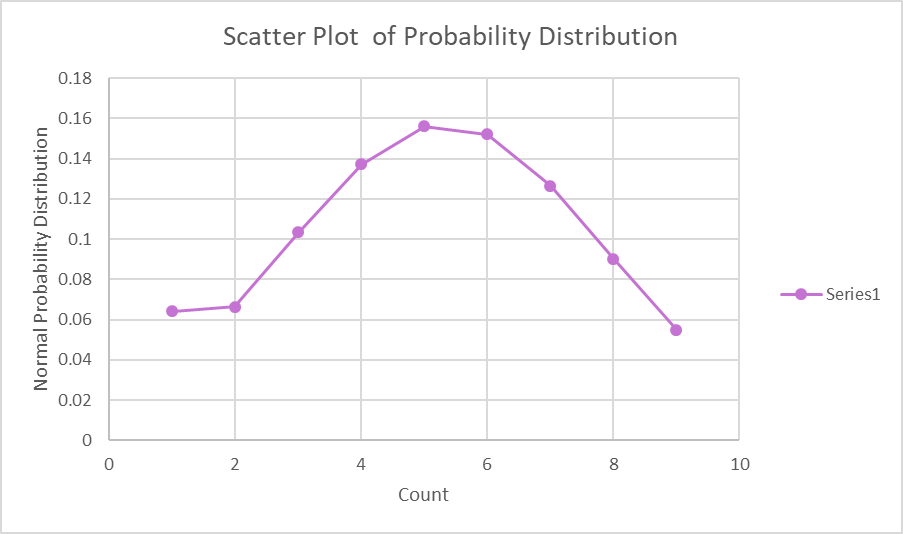
M, N and W table is derived by assigning figures by using the information- M is the number of iterations required to process until an accepted Y value is found, N is the number of accepted values, and W= M/N. After doing all these, we have to work the descriptive statistics, bins, left-right tails, frequencies, relative frequencies only 10 times as Excel gives the count of bins as 10.

Here also my choice of probability distribution will be Normal Probability Distribution only difference being the additional parameter mean in the formula **NORM.DIST(Left/Right tail, mean, SD, 1)**. The first value takes only right tail, from the next onwards difference of right and left tail are taken as X. The expected frequency is multiplying 100 to the normal probability. Ultimately, Chi-square values are measured with our same formula

**(Observed-Expected) ^2 / Expected**.

Chi-Square Goodness of the fit test calculations:

The test statistic is found by summing up all the chi-square values (27.17258). Level of significance is given as 0.05 in the question. Degree of freedom is 94-2-1 i.e., df=91. P-value is obtained using **1-CHISQ.DIST(t-stat, df, 1)**-> 1 is greater than 0.05.



Null Hypothesis: Data is normally distributed

Alternative Hypothesis: Data is not normally distributed

Since our P-value is greater than 0.05, we fail to reject **Ho** and our data is normally distributed. The graph is a proof for the same.

**Conclusion**

This assignment helped in gaining much knowledge about the use of descriptive statistics in Data Analysis. Random number generation is a method which applies to debugging done by simulations and stands as a well-tested generator for encryption algorithms. By solving problem 1 and 2, I understood how important and easy the analysis is using random numbers. Problems 3 and 4 were similar processes. The difference lies in the calculations of normal probabilities- standard deviation in Part 3, standard deviation and mean in Part 4. Including the mean as parameter brings a difference to the shape and skewness of the normal curve. Overall the exhausting work of Excel gave us vision over various kinds of probability distributions, derivation of statistical graphs like histograms, exponential, gamma and normal probability curves and make meaningful observations out of those and validating with the Chi-square Goodness of Fit test which was the emphasis of the whole project.

**Summary**

1. If **r** is a standard uniform random variable, then −𝑳**n**(𝒓) has the **continuous/exponential** probability distribution.
2. The sum of three independent and identically distributed **exponential** random variables has 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 **standard uniform** are used to generate a random value 𝒀 that has the **geometric** probability distribution.

1. The random value 𝑾 that was discussed in problem 4, has the **binomial** probability distribution. The expected value of 𝑾 is: **1.316243** .

**References**

[1] “Probability Distributions in Statistics.” R, 28 Dec. 2015, [www.r-bloggers.com/distributions/](http://www.r-bloggers.com/distributions/).

[2] https://dictionary.apa.org/chi-square-test