Introduction to Enterprise Analytics

Northeastern University

ALY6050, WINTER 2019

MODULE 1 PROJECT ASSIGNMENT

Chi-squared Goodness of Fit Test Project

SUBMITTED BY: SHIVANI ADSAR

NUID: 001399374

SUBMITTED TO: PROF. RICHARD HE

DATE: 01/13/2019

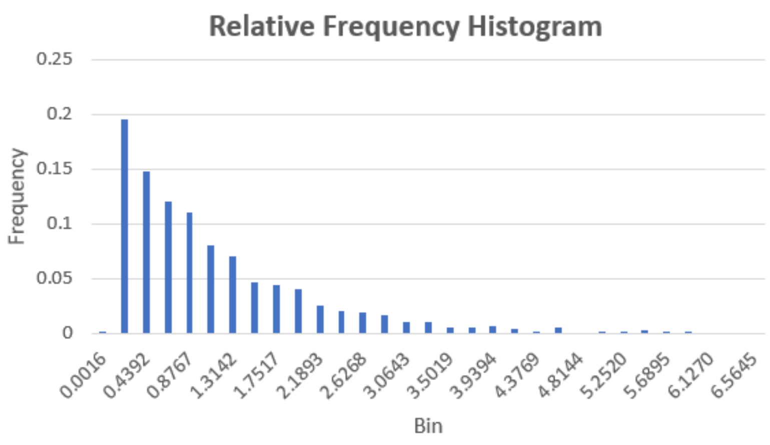
Introduction

The assignment provides practical experience in performing statistical analysis using the randomly generated values. We have performed hypothesis testing to carry out the chi-square test for determining the best fit for the data. Moreover, we have used the normal probability distribution functions to perform analysis and carry out calculations to perform hypothesis testing which would determine the best fit for the data.

Analysis

Part I

1. We have generated random values and calculated the logarithmic values for them by using the logarithmic function. The formula used for performing the logarithmic calculation in excel is =**LN** (value).
2. On the basis of the values of Log, mean and standard deviation, we have calculated the normal probability values using the NORMDIST(value,mean,standard deviation) function.
3. The normal probability values have been used to calculate the bin and frequencies.
   We have used, the formula, Frequency / Total Number of Frequencies, for calculating the relative frequencies.



Observations: The relative frequency histogram shows the frequencies of values with the increase in number of bins.

Fig.1: Relative Frequency Histogram

4. Based on the bin values, we have calculated the left end and right end values which are used to analyses the extreme values used in plotting the bins for the relative frequency histogram.
5. According to the data, it seems that the exponential probability plot is the best fit for the value of X, as the values are logarithmic values.

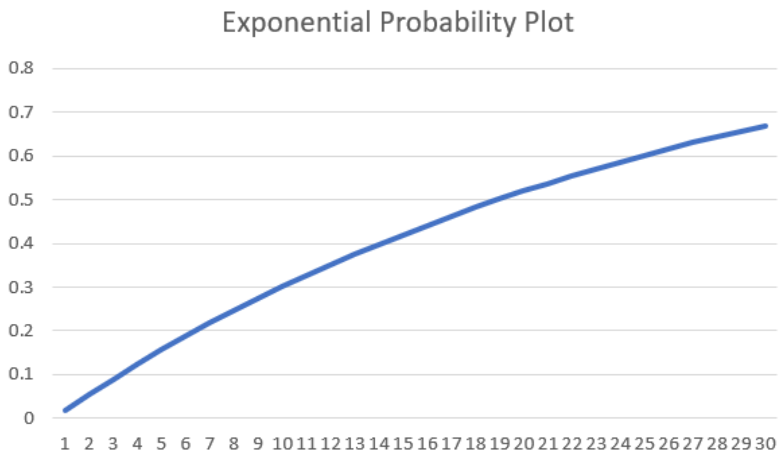


Fig.2: Exponential Probability Plot

Observation: We have used the exponential probability distribution, as it helps us to model the elapsed time.

6. We have performed the chi-squared test of best fit, by calculating the observed frequency and expected frequency, and used the below formula:
   $\sum$(Observed Value-Expected Value)^2 /Expected Value
7. Considering, Ho: Data to be exponentially distributed
   Ha: Data not exponentially distributed
   On performing the chi-square test, it was observed that the p-value ie. 0 is lesser than the significance level and hence we will reject the null hypothesis.

Part II

1. We have generated random values for three columns as R1,R2 and R3. Using a logarithmic function as -LN(R1*R2*R3), we have calculated Rand x.
2. On the basis of Rand x, we have plotted the Relative frequency histogram for the values using the bins and frequencies.

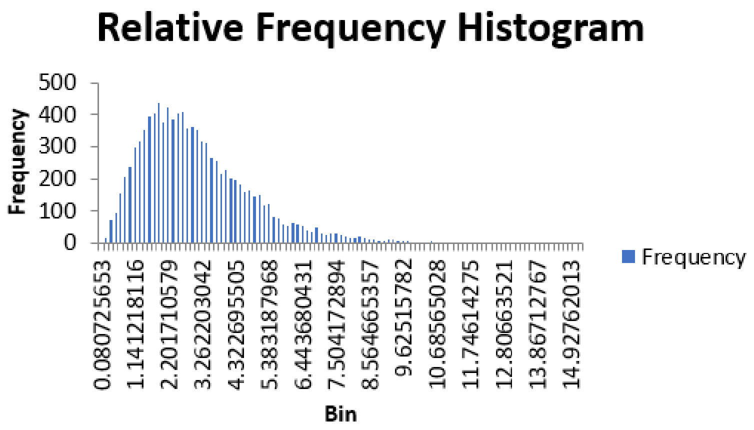


Fig.3: Relative Frequency Histogram

3. Further, we have calculated the Left end and Right end values which have denoted the points on the bins. The observed and expected frequencies have been calculated which have helped in calculating the expected probability based on function,

GAMMA.DIST(Right End,3,1,1) – GAMMA(Left End,3,1,1)

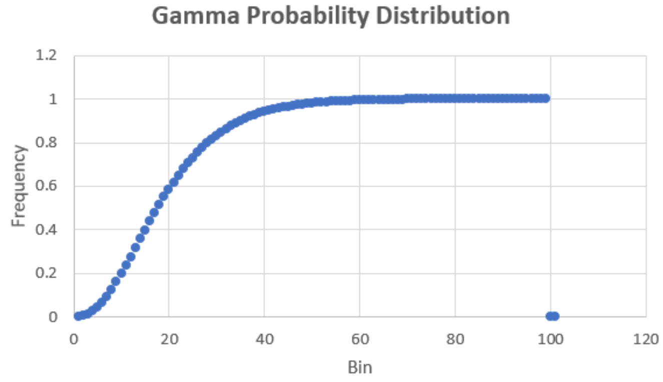


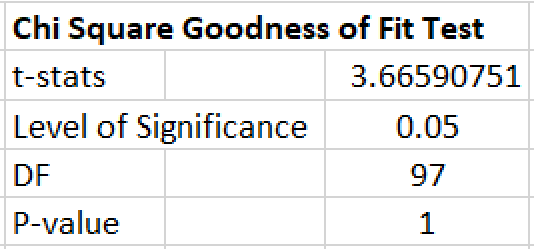
Fig.4: Gamma Probability Distribution

Observations: It can be interpreted, that the frequency becomes constant with the increase in bins for the gamma probability distribution.

4. The expected frequency is calculated based on the expected probability. We have used the gamma function since it converts the factor to non-integer values and generate a smooth curve. We have calculated the chi-square using the formula,

   ∑(Observed Value-Expected Value)^2 /Expected Value

5.



Chi Square Goodness of Fit Test		
t-stats		3.66590751
Level of Significance		0.05
DF		97
P-value		1

Fig.5: Chi-Square Goodness Fit Test

Observations: Considering, Ho: Data displays gamma distribution
Ha: Data does not display gamma distribution
It can be noted that, the p-value is lesser than the significance level, hence we will reject the null hypothesis. It can be interpreted that the data does not show the gamma distribution.

Part III

1. We have generated random values r1 and r2, calculated the logarithm of r1 and r2 for 1000 times, using the function, -LN(value) and stored the result in X1 and X2.
2. Further, the value of K has been calculated using the formula, (X1-1)^2 ÷ 2,
   Using the condition, IF(OR(X1>K,X1=K),RAND,”FALSE”), which means, if X1 > K, it will generate a random number, else, the algorithm will return False. Moreover, we have generated Y values based on, IF((IF X2>=K)=”FALSE”,”NOT RESULT”,IF((IF X2>=K,X1,-X1)), which means if X2 is greater than K, it will display False, else, it will display, NOT RESULT. Also, if r>0.5, we will accept X1 as Y or else we will accept -X1 as Y. In case, if X2<K, the algorithm will return to the initial step 1.
3. Using the Y values, we have plotted the descriptive statistics for min,max,count,mean,range,variance and standard deviation. Using the Y value, we have calculated the bin and frequency to plot the relative frequency histogram.
4. The left end and right end which display the extreme values of a bin, were calculated using the bin values. The relative frequency was calculated as,
   Frequency / Total Number of Frequencies

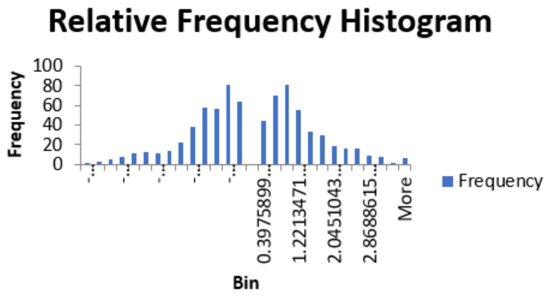


Fig.5: Relative Frequency Histogram

5. We have calculated the Normal Probability plot using, the formula: NORM.DIST(Right End,0,SD,1)-NORM.DIST(Left End,0,SD,1).

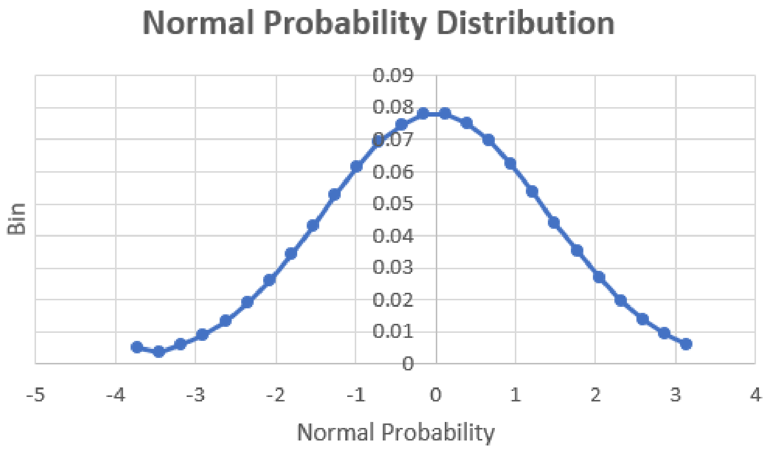
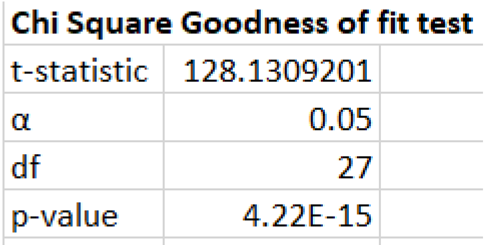


Fig.6: Normal Probability Distribution

6. In this problem, we have calculated the expected frequency, using, count*normal probability. Moreover, using the formula, ∑(Observed Value-Expected Value)^2 /Expected Value, we have calculated the chi-square values and performed the hypothesis testing on the same.



Chi Square Goodness of fit test	
t-statistic	128.1309201
α	0.05
df	27
p-value	4.22E-15

Fig.7: Chi-Square Goodness Fit Test

Observations: Considering,
Ho: Data is normally distributed
Ha: Data is not normally distributed
It can be interpreted, the p-value is greater than significance level, hence we fail to reject the null hypothesis and the data is normally distributed.

Part IV

1. We have used the randomly generated values, R1 and R2 to calculate the logarithmic functions X1 and X2. Further, we have calculated, K using (X1-1)^2/2.
2. We have generated the values of N for every iteration of M, these values denote the number of values without the NA values. Also, W is calculated using M÷N, and used the W values to calculate the descriptive statistics.

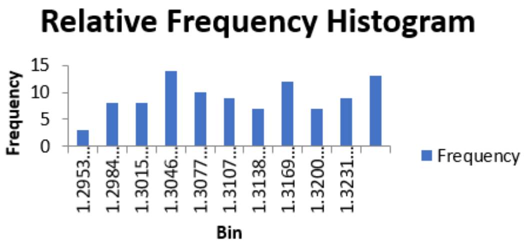


Fig.8: Relative Frequency Histogram

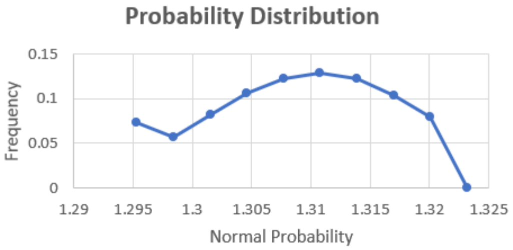
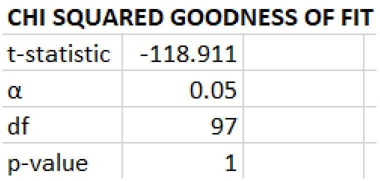


Fig.9: Normal Probability Distribution

3. Further, we have generated the relative frequency histogram, and calculated the Right end, left end, normal probability, expected frequency and chi-square values.



CHI SQUARED GOODNESS OF FIT		
t-statistic	-118.911	
α	0.05	
df	97	
p-value	1	

Fig.10: Chi Square Test

Observation: Considering, Ho: Data normally distributed
Ha: Data is not normally distributed
It can be observed that, the p-value is greater than the significance level, hence we fail to reject the null hypothesis.

Conclusion

1. $-\boldsymbol{L}n(\boldsymbol{r})$ shows the Exponential Probability Distribution, when r is a uniform random variable.
2. The sum of three independent and identically distributed Standard Uniform random variables has the Gamma probability distribution.
3. The problem 3 has an output which displays a normal probability distribution.
4. The random variables X1 and X2 as calculated in the step 2 of problem 3, have probability distribution which is exponential, the random value Y, generated has a normal probability distribution.
5. The random value W, in problem 4 has a normal probability distribution. The expected value is 1.30.

Reference

1. Illowsky, B. (n.d.). Introduction to Statistics. Retrieved from https://courses.lumenlearning.com/introstats1/chapter/the-exponential-distribution/.
2. A. K. (2019, December 22). Gamma Function-Intuition, Derivation, and Examples. Retrieved from https://towardsdatascience.com/gamma-function-intuition-derivation-and-examples-5e5f72517dee.