Northeastern University
College *of* Professional Studies

Chi-squared Goodness of Fit Test Project

Overview and Rationale

This assignment is designed to provide you with hands-on experience in generating random values and performing statistical analysis on those values.

Course Outcomes

This assignment is directly linked to the following key learning outcomes from the course syllabus:

- Use descriptive, Heuristic and prescriptive analysis to drive business strategies and actions

Assignment Summary

Follow the instructions in this project document to generate a number of different random values using random number generation algorithm in Excel, the Inverse Transform. Then apply the Chi-squared Goodness of Fit test to verify whether their generated values belong to a particular probability distribution. Finally, complete a report summarizing the results in your Excel workbook. Submit both the report and the Excel workbook.

The Excel workbook contains all statistical work. The report should explain the experiments and their respective conclusions, and additional information as indicated in each problem. Be sure to include all your findings along with important statistical issues.

Format & Guidelines

The report should follow the following format:

(i) Introduction

(ii) Analysis

(iii) Conclusion

And be 1000 - 1200 words in length and presented in the APA format

Northeastern University
College of Professional Studies

Project Instructions:

The project consists of 4 problems and a summary set of questions. For each problem, tom hints and theoretical background is provided.

Complete each section in a separate worksheet of the same workbook (Excel file). Name your Excel workbook as follows:

ALY6050-Module 1 Project – Your Last Name – First Initial.xlsx

In the following set of problems, $\boldsymbol{r}$ is the standard uniform random value (a continuous random value between 0 and 1).

Problem 1

Generate 1000 random values r. For each r generated, calculate the random value $\boldsymbol{X}$ by:

$\boldsymbol{x = -Ln(r)}$,

where "$\boldsymbol{Ln}$" is the natural logarithm function.

Investigate the probability distribution of $\boldsymbol{X}$ by doing the following:

1. Create a relative frequency histogram of $\boldsymbol{X}$.
2. Select a probability distribution that, in your judgement, is the best fit for $\boldsymbol{X}$.
3. Support your assertion above by creating a probability plot for $\boldsymbol{X}$.
4. Support your assertion above by performing a Chi-squared test of best fit with a 0.05 level of significance.
5. In the word document, describe your methodologies and conclusions.
6. In the word document, explain what you have learned from this experiment.

Hints and Theoretical Background

A popular method for generating random values according to a certain probability distribution is to use the inverse transform method. In this method, the cumulative function of the distribution (*F(x)*) is used for such a random number generation. More specifically, a standard uniform random value *r* is generated first. Most software environments are capable of generating such a value. In Excel and *R*, functions "=RAND()" and "runif()" generate such a value respectively. After *r* has been created, it then replaces *F(x)* in the expression of the cumulative function and the resulting equation is solved for the variable *x*.

For example, suppose we wish to generate a random value according to the exponential distribution with a certain mean (say μ). The cumulative function for the exponential distribution is:

$$\boldsymbol{F(x) = 1 - e^{-\frac{1}{\mu}x}}$$

(The quantity *1/μ* in the above description is called the rate of the exponential random variable and is denoted by λ.)

Therefore, to generate a random value x that belongs to the exponential distribution with a mean of μ. We first generate a standard uniform value r, then replace *F(x)* by *r* in the above expression, and solve the resulting equation for the variable *x*:

$$\boldsymbol{r = 1 - e^{-\frac{1}{\mu}x}}$$

$$\boldsymbol{e^{-\frac{1}{\mu}x} = 1 - r}$$

$$\boldsymbol{-\frac{1}{\mu}\, x = Ln(1 - r)}$$

$$\boldsymbol{x = -\mu\, Ln(1 - r)}$$

The formula above means that if *R* is a standard uniform random variable, then the random variable *X* obtained by the expression $\boldsymbol{X = -\mu\, Ln(1 - R)}$ will belong to the exponential distribution with an average which is equal to the value of μ. This formula can be simplified as:

$$\boldsymbol{X = -\mu\, Ln(R)}$$

(Note that If *R* is a standard uniform random variable, then $(1 - \boldsymbol{R})$ is also standard uniform.)

A special case of the above formula is when $\boldsymbol{\mu = 1}$. This means that a random variable *x* generated by the formula $\boldsymbol{X = -Ln(R)}$is an exponential random variable with an average of 1 (or, rate=1).

Problem 2

Generate three sets of standard uniform random values, $\boldsymbol{r_1}$, $\boldsymbol{r_2}$and $\boldsymbol{r_3}$**,** each consisting of 10,000 values. Next, calculate the random value $\boldsymbol{x}$ according to the following formula:

$$\boldsymbol{x = -Ln(r_1 r_2 r_3).}$$

Investigate the probability distribution of $\boldsymbol{X}$ by doing the following:

1. Create a relative frequency histogram of $\boldsymbol{X}$.
2. Select a probability distribution that, in your judgement, is the best fit for $\boldsymbol{X}$.
3. Support your assertion above by creating a probability plot for $\boldsymbol{X}$.
4. Support your assertion above by performing a Chi-squared test of best fit with a 0.05 level of significance.
5. In the word document, describe your methodologies and conclusions.
6. In the word document, explain what you have learned from this experiment.

Hints and Theoretical Background:

This problem is related to a theorem in the probability theory. The theorem states that:

If $\boldsymbol{X_1, X_2, \ldots, X_n}$are $\boldsymbol{n}$ identical and independent exponential random variables each with a mean of μ, then the random variable obtained by their sum, that is $\boldsymbol{X_1 + X_2 + \ldots + X_n}$, will have a $\boldsymbol{Gamma(n, \mu)}$ probability distribution, where n is the ***shape parameter*** of the Gamma distribution and $\boldsymbol{\mu} = \frac{\boldsymbol{1}}{\boldsymbol{rate}}$.

From the Hints and Theoretical Background of Problem 1, we know that if R is a standard uniform random variable, then $\boldsymbol{X} = -\boldsymbol{Ln}(\boldsymbol{R})$ is an exponential random variable with an average of 1. Therefore, if $\boldsymbol{R_1}$, $\boldsymbol{R_2}$**,** and $\boldsymbol{R_3}$ are three independent standard uniform random variables, then $\boldsymbol{X_1} = -\boldsymbol{Ln}(\boldsymbol{R_1}))$, $\boldsymbol{X_2} = -\boldsymbol{Ln}(\boldsymbol{R_2})$, and $\boldsymbol{X_3} = -\boldsymbol{Ln}(\boldsymbol{R_3})$are three independent and identical (each with a mean of 1) exponential random variables. Thus, according to the theorem above, the random variable formed by their sum, that is $(-\boldsymbol{Ln}(\boldsymbol{R_1})) + (-\boldsymbol{Ln}(\boldsymbol{R_2})) + (-\boldsymbol{Ln}(\boldsymbol{R_3}))$, will belong to the $\boldsymbol{Gamma}(\boldsymbol{3}, \boldsymbol{1})$probability distribution.

However algebraically,

$$(-\boldsymbol{Ln}(\boldsymbol{R_1})) + (-\boldsymbol{Ln}(\boldsymbol{R_2})) + (-\boldsymbol{Ln}(\boldsymbol{R_3})) = -(\boldsymbol{Ln}(\boldsymbol{R_1}) + \boldsymbol{Ln}(\boldsymbol{R_2}) + \boldsymbol{Ln}(\boldsymbol{R_3})) = -\boldsymbol{Ln}(\ \boldsymbol{R_1 R_2 R_3}).$$

Therefore, if $\boldsymbol{R_1}$, $\boldsymbol{R_2}$**,** and $\boldsymbol{R_3}$ are three independent standard uniform random variables between zero and 1, then the random variable X formed by the formula $\boldsymbol{X} = -\boldsymbol{Ln}(\ \boldsymbol{R_1 R_2 R_3})$ will belong to the $\boldsymbol{Gamma}(\boldsymbol{3}, \boldsymbol{1})$ probability distribution.

Problem 3

Generate a set of 1000 pairs of standard uniform random values $r_1$and r_2. Then perform the following algorithm for each of these 1000 pairs: Let the output of this algorithm be denoted by Y.

Step 1: Generate random values $X_1 = -Ln(r_1)$ and $X_2 = -Ln(r_2)$

Step 2: Calculate $k = \frac{(x_1-1)^2}{2}$. If $x_2 \geq k$, then generate a random number r. If $r > 0.5$ accept $x_1$as Y(that is, let $Y = x_1$); otherwise if $r \leq 0.5$, else accept $-x_1$ as Y (that is, let $Y = -x_1$).

If $x_2 < k$, no result is obtained, and the algorithm returns to step 1. This means that the algorithm skips the pair r_1 and r_2 for which $x_2 < K$ without generating any result and moves to the next pair r_1 andr_2.

After repeating the above algorithm 1000 times, a number *N* of the *Y* values will be generated. Obviously $N \leq 10,000$ since there will be instances when a pair r_1 and r_2 would not generate any result, and consequently that pair would be wasted.

Investigate the probability distribution of Y by doing the following:

1. Create a relative frequency histogram of Y.
2. Select a probability distribution that, in your judgement, is the best fit for Y.
3. Support your assertion above by creating a probability plot for Y.
4. Support your assertion above by performing a Chi-squared test of best fit with a 0.05 level of significance.
5. In the word document, describe your methodologies and conclusions.
6. In the word document, explain what you have learned from this experiment.

Hints and Theoretical Background

Other than the inverse transform method used for generating random values that are according to a certain particular probability distribution, a second applied method for generating random values is the Rejection algorithm. The details of this algorithm are explained below:

Suppose we wish to generate random values *x* that is according to a certain probability distribution with $f(x)$as its probability density function (*pdf*). Also suppose that the following two conditions are satisfied

(i) we are able to generate random values *y* that belong to a probability distribution whose probability density function is $\boldsymbol{g(y)}$,

(ii) there exists a positive constant *C* such that $\frac{f(y)}{g(y)} \le \boldsymbol{C}$ for all *y* values (this means that the ratio $(\frac{f(y)}{g(y)})$ is always bounded and does not grow indefinitely. This condition is almost always satisfied for any two probability density functions $\boldsymbol{f(x)}$ and $\boldsymbol{g(y)}$).

The rejection algorithm can now be implemented as follows:

Step 1: Generate a random value y that belongs to the probability distribution with $\boldsymbol{g(y)}$ as its pdf and generate a standard uniform random value *r*.

Step 2: Evaluate $\boldsymbol{k} = \frac{f(y)}{C\,g(y)}$. *If* $\boldsymbol{r} \le \boldsymbol{k}$, then accept *y* as the random variable *x* (that is, let $\boldsymbol{x} = \boldsymbol{y}$); otherwise return to Step1 and try another pair of $(\boldsymbol{y}\,, \boldsymbol{r})$ values.

A few remarks about the Rejection algorithm is worth noting:

1. The probability that the generated *y* value will be accepted as *x*, is: $\frac{f(y)}{C\,g(y)}$. This is the reason why the algorithm uses a standard uniform value *r* and accepts *y* as *x* if $\boldsymbol{r} \le \frac{f(y)}{C\,g(y)}$.
2. Each iteration of the algorithm will independently result in an accepted value with a probability equal to: $\boldsymbol{P}\left(\boldsymbol{r} \le \frac{f(y)}{C\,g(y)}\right) = \frac{1}{C}$. Therefore, the number of iterations needed to generate one accepted *y* value follows a geometric probability distribution with mean *C*.

Relevancy of Problem 3 to the Rejection Algorithms:

In problem 3, the random variable y , selected from an exponential probability distribution with *rate* =1 and a pdf of $\boldsymbol{g(y)} = \boldsymbol{e^{-y}}$, is used to first generate the absolute value of a standard normal random variable *x* ($|\boldsymbol{x}|$has the pdf: $\boldsymbol{f(x)} = \frac{2}{\sqrt{2\pi}}\,\boldsymbol{e}^{-\frac{x^2}{2}}$), and then assign positive or negative signs to this value (through a standard uniform variable *r*) in order to obtain a standard normal random value. It can be shown algebraically that $\frac{f(y)}{g(y)} = \sqrt{\frac{2e}{\pi}}\;\boldsymbol{e}^{\frac{-(y-1)^2}{2}} \le \sqrt{\frac{2e}{\pi}}$ for all *y* values (note that $\boldsymbol{e}^{\frac{-(y-1)^2}{2}} \le \boldsymbol{1}$ for all *y* values). Therefore, the constant C in the assumptions of the algorithm can be chosen to be: $\boldsymbol{C} = \sqrt{\frac{\boldsymbol{2e}}{\boldsymbol{\pi}}} \approx$ $\mathbf{1.315}$. Therefore, $\frac{\boldsymbol{f}(y)}{\boldsymbol{C}\,\boldsymbol{g}(y)} = \boldsymbol{e}^{\frac{-(y-1)^2}{2}}$. Hence the following algorithm can be used to generate the absolute value of a standard normal random variable:

Step 1: Generate random variables $\boldsymbol{Y}$ and $\boldsymbol{R}$; with $\boldsymbol{Y}$ being exponential with *rate=1*, and $\boldsymbol{R}$ being uniform on $(\mathbf{0}, \mathbf{1})$

Step 2: If $\boldsymbol{R} \le \boldsymbol{e}^{\frac{-(y-1)^2}{2}}$, then accept $\boldsymbol{Y}$ as the random variable $\boldsymbol{X}$ (that is, set $\boldsymbol{X} = \boldsymbol{Y}$); otherwise return to Step1 and try another pair of $(\boldsymbol{Y}\,, \boldsymbol{R})$ values.

Note that in step 2 of the above algorithm, the condition $\boldsymbol{R} \le \boldsymbol{e}^{\frac{-(y-1)^2}{2}}$ is mathematically equivalent to: $-\boldsymbol{Ln}(\boldsymbol{R}) \le \frac{(y-1)^2}{2}$. However, we have already seen in the Hints and Theoretical Backgrounds of the earlier problems that if $\boldsymbol{R}$ is standard uniform, then $-\boldsymbol{Ln}(\boldsymbol{R})$ is exponential with *rate=1*. Therefore, the algorithm for generating the absolute value of the standard normal random variable can be modified as follows:

Step 1: Generate independent exponential random variables $\boldsymbol{Y_1}$ and $\boldsymbol{Y_2}$; each with *r*ate=1.

Step 2: Evaluate $\boldsymbol{k} = \frac{(y_1-1)^2}{2}$2. If $\boldsymbol{k} \le \boldsymbol{Y_2}$, then accept $\boldsymbol{Y_1}$ as the random variable $\boldsymbol{X}$ (that is, set $\boldsymbol{X} = \boldsymbol{Y_1}$); otherwise return to Step1 and try another pair of $(\boldsymbol{Y_1}\,, \boldsymbol{Y_2})$ values.

In fact, it is the above version of the Rejection algorithm that is being implemented in Problem 3. However, in order to obtain a standard normal random value (instead of its absolute value), the step 2 of the above algorithm has been modified as follows:

Step 2: Evaluate $\boldsymbol{k} = \frac{(y_1-1)^2}{2}$. If $\boldsymbol{k} \le \boldsymbol{Y_2}$, then generate a standard uniform variable $\boldsymbol{R}$. If $\boldsymbol{R} \ge \mathbf{0.5}$, set $\boldsymbol{X} = \boldsymbol{Y_1}$, otherwise set $\boldsymbol{X} = -\boldsymbol{Y_1}$. If $\boldsymbol{k} > \boldsymbol{Y_2}$, return to step 1 and try another pair of $(\boldsymbol{Y_1}\,, \boldsymbol{Y_2})$ values.

Note: The standard normal random value generated by the Rejection algorithm can be used to generate any normal random value with a mean μ and a standard deviation σ. Once a standard normal variable Z has been generated, it suffices to evaluate $\boldsymbol{\mu} + \boldsymbol{\sigma Z}$ to generate the desired normal variable.

Problem 4

In the algorithm of problem #3 above, there are instances when the generated random values do not satisfy the condition $\boldsymbol{x_2} \geq \boldsymbol{k}$ In order to obtain an acceptable value for $\boldsymbol{Y}$. In such cases, the algorithm returns to *step 1* and generates another two values to check for acceptance. Let $\boldsymbol{M}$ be the number of iterations needed to generate $\boldsymbol{N}$ of the accepted $\boldsymbol{Y}$ values $(\boldsymbol{M} \geq \boldsymbol{N})$. Let $\boldsymbol{W} = \frac{\boldsymbol{M}}{\boldsymbol{N}}$**.**

(For example, suppose that the algorithm has produced 700 $\boldsymbol{Y}$ values $(\boldsymbol{N} = \mathbf{700})$ after 1000 iterations $(\boldsymbol{M} = \mathbf{1000})$. Then $\boldsymbol{W} = \frac{\mathbf{1000}}{\mathbf{700}} = \mathbf{1.43}$. This means that it takes the algorithm 1.43 iterations to produce one output. In fact, $\boldsymbol{W}$ itself is a random variable. Theoretically, $\boldsymbol{E}(\boldsymbol{W})$ - the expected value (i.e., average) of $\boldsymbol{W}$ – of an algorithm is a measure of efficiency of that algorithm.)

Investigate $\boldsymbol{W}$ by the following sequence of exploratory data analytic methods:

1. Estimate the ***expected value*** and the ***standard deviation*** of $\boldsymbol{W}$.
2. Select a probability distribution that, in your judgement, is the best fit for $\boldsymbol{W}$.
3. Support your assertion above by performing a Chi-squared test of best fit with a 0.05 level of significance**.**
4. As the number of iterations $\boldsymbol{M}$ becomes larger, the values $\boldsymbol{W}$ will approach a certain limiting value. Investigate this limiting value of $\boldsymbol{W}$ by completing the following table and plotting $\boldsymbol{W}$ versus $\boldsymbol{M}$. What value do you propose for the limiting value that $\boldsymbol{W}$ approaches to?

M	W
10	
20	
30	
40	
50	
60	
70	

80	
90	
100	
200	
300	
400	
500	
600	
700	
800	
900	
1000	

5. In the word document, communicate to the reader your findings about $\boldsymbol{W}$.
6. In the word document, explain what you have learned from this experiment.

Summary

In the word document, summarize and conceptualize your findings in parts 1 – 4 above by filling the blanks in the sentences below:

1. If $\boldsymbol{r}$ is a standard uniform random variable, then $-\boldsymbol{Ln}(\boldsymbol{r})$ has the ________________ probability distribution.

2. The sum of three independent and identically distributed _______________ random variables has the _____________________ probability distribution.

3. The output of the algorithm of problem 3 has a ___________________ probability distribution.

4. In step 2 of the algorithm of problem 3, random variables $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$, each of whose probability distribution is _____________________ are used to generate a random value $\boldsymbol{Y}$ that has the ____________________probability distribution.

5. The random value $\boldsymbol{W}$ that was discussed in problem 4, has the ____________________________ probability distribution. The expected value of $\boldsymbol{W}$ is: __________.