Mustafa Abdullayev

Random Numbers and Chi-Square Goodness of Fit

AY6050 – Intro to Enterprise Analytics

Term: Winter 2020

Instructor: Roy Wada

02/24/2020

# Introduction

This is Microsoft Word Report accompanying R Script. In my Script, my main aim was to use descriptive, heuristic and prescriptive analysis to drive business actions. As a data source, I utilized the Random values on R and performed statistical analysis using R. I generated the different random values and used Chi-Square Goodness of Fit Test to verify whether the generated values belong to a particular probability distribution. Also, I utilized powerful R built-in functions and graphs, such as bar charts, to dive deeper to observe hidden patterns and visually communicate my findings to the audience. Since I also provided R script with all the codes and comments, I removed some of the codes and comments from my report (such as package loading). It is due to keep my report brief, succinct and to the point.

# Problem 1

For the first problem I generated 1000 uniformly distributed random values between 0 and 1. These random values are generated by using inverse transform method. Then I created the new vector ,X, by calculating the Natural Logarithm of these values. I was interested with the probability distribution of X values. In order to find that, I created the Relative Frequency table for X values (Figure 1.1). From this figure, it seems the X values are very likely to fit Exponential distribution. In order to dive deeper, I created the Probability Plot for X. (Figure 1.2). I plotted X values vs Randomly generated Exponentially distributed values and observed that there is an almost perfect fit. So , I claimed that X values, which are generated from calculating the negative Natural Logarithm of uniformly distributed values, has the Exponential Distribution. In order to prove my claim, I conducted Chi-Square Goodness of Fit test with 0.05 significance. Here my hypotheses are as follow :

H0 : There is no significant difference between observed (X) and expected values(Exp. distribution)

H1 : There is a significant difference between observed(X) and expected values (Exp. distribution)

With the help of R, I calculated the p-value for this test is 0.9404 and it was smaller than my significance level. So, I am unable to reject the Null hypothesis. Therefore, I have a strong evidence suggesting with the 95% confidence that, X values, which were generated from taking negative logarithm of uniformly distributed random values, has the shape of Exponential Probability Distribution.

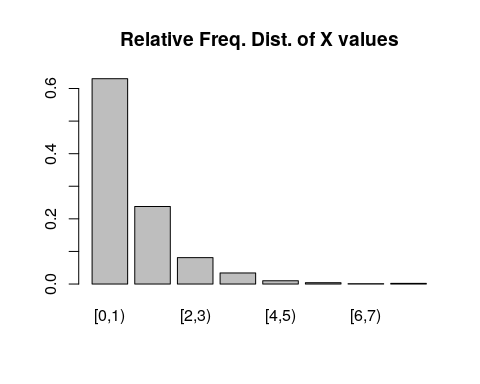
##################### Problem 1 ####################

## creating 1000 random numbers between 0 and 1  
r <- runif(1000, min = 0, max = 1)  
## Getting negative logarithmic values for r  
x\_1 <- -log(r)  
  
## I will use bins for these values - x  
bins\_1 <- seq(0,8,by = 1)  
  
## seperating x vector into 8 bins  
x\_1.cut <- cut(x\_1,bins\_1,right = FALSE)

## making frequency table by combining x.cut into table  
x\_1.freq <- table(x\_1.cut)  
barplot(x\_1.freq, main = "Frequency Distribution of X values")

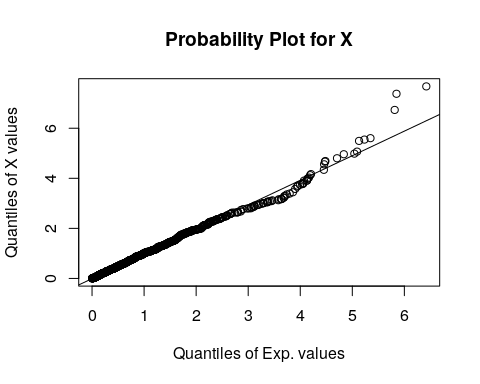
## Calculating Relative frequency by dividing 1000 (total numbers)  
x\_1.r\_freq <- x\_1.freq/1000  
barplot(x\_1.r\_freq, main = "Relative Freq. Dist. of X values")

Figure .1 – Relative Frequency Distribution of X



## Creating Probability plot  
y\_1 <- rexp(1000)  
qqPlot(y\_1,x\_1,plot.type = "Q-Q",add.line = TRUE,main = "Probability Plot for X",xlab = "Quantiles of Exp. values",ylab = "Quantiles of X values")

Figure 1.2 – Probability Plot for X



##Goodness of fit test :   
## Here : H0 : There is no significant difference between the observed and the expected values  
## HA : There is a significant difference between the observed and the expected values  
##Significance level = 0.05  
  
x\_1.exp <- c(pexp(1),pexp(2)-pexp(1),pexp(3)-pexp(2),pexp(4)-pexp(3),pexp(5)-pexp(4),pexp(6)-pexp(5),pexp(7)-pexp(6),1- pexp(7))

res\_1 <- chisq.test(x\_1.freq,p = x\_1.exp)

res\_1

#Chi-squared test for given probabilities  
  
#data: x.freq  
#X-squared = 2.3143, df = 7, **p-value = 0.9404**

# Problem 2

For the second Problem, again with same logic I generated 3 different random number vectors, each one has 10000 values, between 0 and 1 according uniform distribution. Afterward, I for each value in these 3 vectors, I calculated X as the negative natural logarithm of their multiplication (X = -Ln(r1\*r2\*r3)). Again, I was interested in the distribution of X values. This time, I utilized the mathematical theorem stating that sum of n identical and independent exponential distributions is gamma distribution with shape = 3 (“Relationship among probability distributions”,2019). Also, from problem 1 we observed that negative natural logarithm of random values generated according to uniform distribution will have Exponential distributions. Basic Algebra here shows that our X , (which is -Ln(r1\*r2\*r3) = -(Ln(r1) + Ln(r2) + Ln(r3))) is indeed sum of 3 independent and identical Exponential random variables. So, I assume that X has the Gamma Probability distribution with n=3(shape). In order to visualize that, I constructed Relative Frequency Plot (Figure 2.1) and Probability Plot for X values (Figure 2.2). Form Probability plot ,we can observe that there is an almost perfect fit between Gamma (shape = 3) and X values. Also, Relative frequency plot support this claim. So, in order to formally prove this assumptions, I conducted a hypothesis testing. I utilized the Chi – Square Test with the significance of 0.05. Here my assumptions are as follow :

H0 : There is no significant difference between observed (X) and expected values(Gamma. distribution)

H1 : There is a significant difference between observed(X) and expected values (Gamma. distribution)

It turns out that, the p-value for this test is 0.3195 and it was smaller than my significance level. So, I am unable to reject the Null hypothesis. Therefore, I have a strong evidence suggesting with the 95% confidence that, X values, which were generated from taking negative logarithm of 3 different vectors of uniformly distributed random values and summing them , has the shape of Gamma Probability Distribution with n=3.

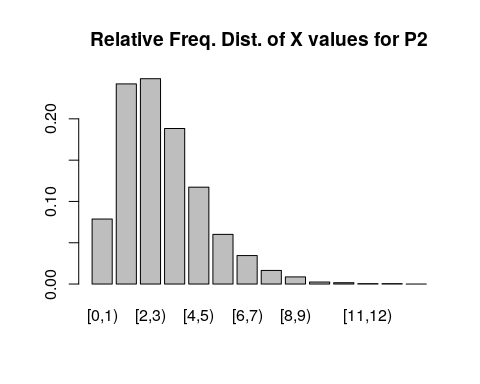
#################### Problem 2 ####################  
  
r1 <- runif(10000, min = 0, max = 1)  
r2 <- runif(10000, min = 0, max = 1)  
r3 <- runif(10000, min = 0, max = 1)  
  
x\_2 <- -log(r1\*r2\*r3)  
bins\_2 <- seq(0,14,by = 1)

x\_2.cut <- cut(x\_2,bins\_2,right = FALSE)

## making frequency table by combining x.cut into table  
x\_2.freq <- table(x\_2.cut)  
barplot(x\_2.freq, main = "Frequency Distribution of X values for P2 ")

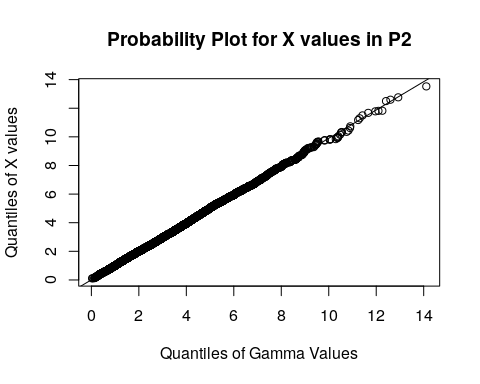
## Calculating Relative frequency by dividing 1000 (total numbers)  
x\_2.r\_freq <- x\_2.freq/10000  
barplot(x\_2.r\_freq, main = "Relative Freq. Dist. of X values for P2")

Figure .1 - Relative Frequency Distribution of X



## Creating Probability plot  
y\_2 <- rgamma(10000,shape = 3)  
  
qqPlot(x\_2,y\_2,add.line = TRUE,main = "Probability Plot for X values in P2",xlab = "Quantiles of Gamma Values",ylab = "Quantiles of X values")

Figure 2.2 - Probability Plot for X



##Goodness of fit test :   
## Here : H0 : There is no significant difference between the observed and the expected values  
## HA : There is a significant difference between the observed and the expected values  
##Significance level = 0.05  
  
exp\_2 <- c(pgamma(shape = 3,1)-pgamma(shape = 3,0),pgamma(shape = 3,2)-pgamma(shape = 3,1),pgamma(shape = 3,3)-pgamma(shape = 3,2),pgamma(shape = 3,4)-pgamma(shape = 3,3),pgamma(shape = 3,5)-pgamma(shape = 3,4),pgamma(shape = 3,6)-pgamma(shape = 3,5),pgamma(shape = 3,7)-pgamma(shape = 3,6),pgamma(shape = 3,8)-pgamma(shape = 3,7),pgamma(shape = 3,9)-pgamma(shape = 3,8),pgamma(shape = 3,10)-pgamma(shape = 3,9),pgamma(shape = 3,11)-pgamma(shape = 3,10),pgamma(shape = 3,12)-pgamma(shape = 3,11),pgamma(shape = 3,13)-pgamma(shape = 3,12),1-pgamma(shape = 3,13)

res\_2 <- chisq.test(x=x\_2.freq,p=exp\_2)

res\_2

#Chi-squared test for given probabilities  
  
#data: x\_2.freq  
#X-squared = 14.807, df = 13**, p-value = 0.3195**

# Problem 3

For the problem 3, I generated the Y values according to the algorithm given in the Problem 3 of [assignment sheet](https://s3.us-east-1.amazonaws.com/blackboard.learn.xythos.prod/5a3148150d016/15711077?response-content-disposition=inline%3B%20filename%2A%3DUTF-8%27%27ALY6050_Project%25201%2520Document_4%2520column%2520rubric%25282%2529.pdf&response-content-type=application%2Fpdf&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Date=20200224T173444Z&X-Amz-SignedHeaders=host&X-Amz-Expires=21600&X-Amz-Credential=AKIAIL7WQYDOOHAZJGWQ%2F20200224%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Signature=03869727095e548d5303f5557733bd9d4c7f3a4c0ee1836b8372f7693755e48b) . Also, for this problem, I was interested in the probability distribution of Y values. Again, I created the Relative Frequency plot (Figure 3.1). It is obvious that distribution of Y values highly resemblances the Standard Normal Distribution. For Further analysis, I created the Probability Plot for Y values vs Standard Normal Values and observed that there is an almost perfect fit. (Figure 3.1). Finally, for the formal proof, I conducted a hypothesis testing. I utilized the Chi – Square Test with the significance of 0.05. Here my assumptions are as follow :

H0 : There is no significant difference between observed (Y) and expected values(Standard Normal distribution)

H1 : There is a significant difference between observed(Y) and expected values (Standard Normal distribution)

It turns out that, the p-value for this test is 0.7257 and it was smaller than my significance level. So, I am unable to reject the Null hypothesis. Therefore, I have a strong evidence suggesting with the 95% confidence that, Y values, which were generated according to the algorithm given in the assignment sheet, has the shape of Standard Normal Probability Distribution.

#################### Problem 3 ####################  
  
r\_3\_1 <- runif(1000)  
r\_3\_2 <- runif(1000)  
  
x\_3\_1 <- -log(r\_3\_1)  
x\_3\_2 <- -log(r\_3\_2)  
  
Y <- vector()

k <- ((x\_3\_1 - 1)^2)/2  
  
for (val in seq(1000)) {  
   
 if(x\_3\_2[val] >= k[val]){  
 r <- runif(1)  
   
 if (r > 0.5){  
 Y <- append(Y,x\_3\_1[val])  
 }  
 else{  
 Y <-append(Y,-x\_3\_1[val])  
 }  
 }  
   
}  
  
N <- length(Y)

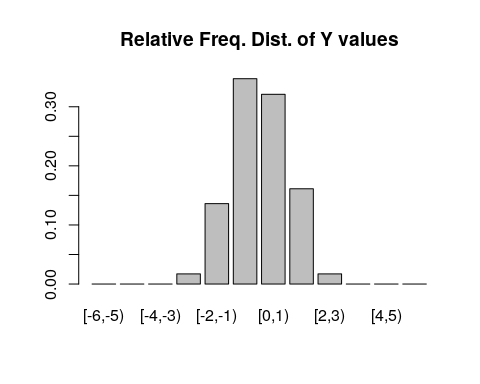
bins\_3 <- seq(-6,6,by = 1)

Y.cut <- cut(Y,bins\_3,right = FALSE)

## making frequency table by combining x.cut into table  
Y.freq <- table(Y.cut)  
barplot(Y.freq, main = "Frequency Distribution of Y values")

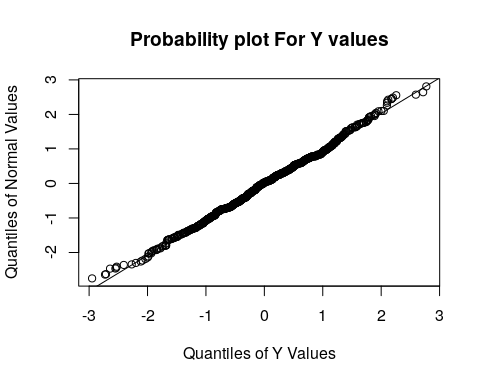
## Calculating Relative frequency by dividing N (total numbers)  
Y.r\_freq <- Y.freq/N  
barplot(Y.r\_freq, main = "Relative Freq. Dist. of Y values")

Figure .1 - Relative Frequency Distribution for Y



## Creating Probability plot  
  
qqPlot(Y,rnorm(N),add.line = TRUE,main = "Probability plot For Y values",xlab = "Quantiles of Y Values",ylab="Quantiles of Normal Values")

Figure 3.2 - Probability Plot for Y



##Goodness of fit test :   
## Here : H0 : There is no significane difference between the observed and the expected values  
## HA : There is a significant difference between the observed and the expected values  
##Significance level = 0.05  
  
norm\_3 <- c(pnorm(-5),pnorm(-4)-pnorm(-5),pnorm(-3)-pnorm(-4),pnorm(-2)-pnorm(-3),pnorm(-1)-pnorm(-2),pnorm(0)-pnorm(-1),pnorm(1)-pnorm(0),pnorm(2)-pnorm(1),pnorm(3)-pnorm(2),pnorm(4)-pnorm(3),pnorm(5)-pnorm(4),1 - pnorm(5))  
sum(norm\_3)

## [1] 1

res\_3 <- chisq.test(x=Y.freq,p=norm\_3)

#Chi-squared test for given probabilities  
  
#data: Y.freq  
#X-squared = 7.8609, df = 11, p-value = 0.7257

# Problem 4

In the problem above, in order to generate N numbers, we had to do given algorithm M times. (obviously M >= N). So, let’s say W = M/N. Then this W value can be thought as an efficiency of our calculation. Since W itself is a random variable, I am interested in its distribution. Using the same procedure above, I realized that W has a similar distribution with Normal Distribution. From its Relative Frequency graph, we can observe this similarity (Figure 4.1). Also, I utilized Probability Plot to see if W values fit the Normal Distribution. (Figure 4.2). Indeed, there is fit. So, in order to formally prove that, I conducted a hypothesis test. I utilized the Chi – Square Test with the significance of 0.05. Here my assumptions are as follow :

H0 : There is no significant difference between observed (W) and expected values(Normal distribution)

H1 : There is a significant difference between observed(W) and expected values (Normal distribution)

It turns out that, the p-value for this test is 0.7044 and it was smaller than my significance level. So, I am unable to reject the Null hypothesis. Therefore, I have a strong evidence suggesting with the 95% confidence that, W values, which were generated according to the algorithm given in the assignment sheet, has the shape of Normal Probability Distribution.

Also, I tried find the limit of W values. I plotted W values vs various values to see what happens to W. From figure 4.3, it suggests that expected values for W is around 1.30. (Limit value). Indeed, after calculating the mean and standard deviation of W, I found that W has 1.30 Expected value and 0.091 Standard Deviation.

#################### Problem 4 ####################  
set.seed(123)  
  
to\_test <- c(10,20,30,40,50,60,70,80,90,100,200,300,400,500,600,700,800,900,1000)  
W <- vector()  
  
## Algorithm Function  
for (num in to\_test){  
 r\_4\_1 <- runif(num)  
 r\_4\_2 <- runif(num)  
  
 x\_4\_1 <- -log(r\_4\_1)  
 x\_4\_2 <- -log(r\_4\_2)  
  
 Y\_2 <- vector()  
 Y\_2  
  
 k\_2 <- ((x\_4\_1 - 1)^2)/2  
  
 for (val in seq(num)) {  
   
 if(x\_4\_2[val] >= k\_2[val]){  
 r <- runif(1)  
   
 if (r > 0.5){  
 Y\_2 <- append(Y\_2,x\_4\_1[val])  
 }  
 else{  
 Y\_2 <-append(Y\_2,-x\_4\_1[val])  
 }  
 }  
 }  
   
 N <- length(Y\_2)  
 W <- append(W,num/N)  
}  
  
  
## Expected value and standard deviation  
mean(W)

## [1] 1.298635

sd(W)

## [1] 0.09057059

## Lets Normalize the W values  
norm\_W <- (W-mean(W))/sd(W)  
  
## Seperating to bins  
bins\_4 <- seq(-5,5,by = 1)  
bins\_4

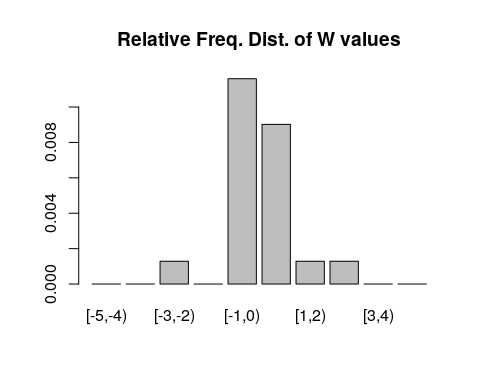
## [1] -5 -4 -3 -2 -1 0 1 2 3 4 5

W.cut <- cut(norm\_W,bins\_4,right = FALSE)

## making frequency table by combining x.cut into table  
W.freq <- table(W.cut)  
barplot(W.freq, main = "Frequency Distribution of W values")

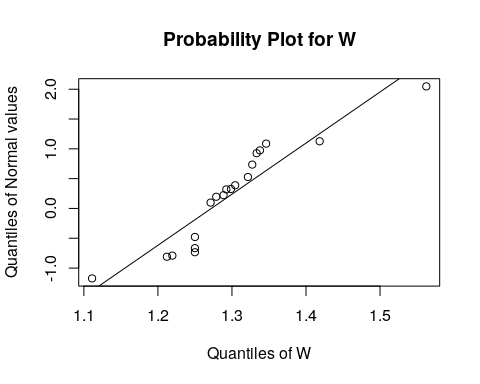
## Calculating Relative frequency by dividing N (total numbers)  
W.r\_freq <- W.freq/N  
barplot(W.r\_freq, main = "Relative Freq. Dist. of W values")

Figure .1 - Relative Frequency Distribution of W



## Probability plot for W values  
  
qqPlot(W,rnorm(length(W)),add.line = TRUE,main = "Probability Plot for W",ylab = "Quantiles of Normal values")

Figure 4.2 - Probability Plot for W

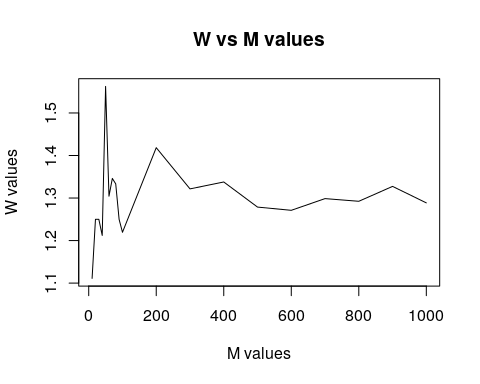


##Goodness of fit test :   
## Here : H0 : There is no significane difference between the observed and the expected values  
## HA : There is a significant difference between the observed and the expected values  
##Significance level = 0.5  
  
W\_norm <- c(pnorm(-4),pnorm(-3)-pnorm(-4),pnorm(-2)-pnorm(-3),pnorm(-1)-pnorm(-2),pnorm(0)-pnorm(-1),pnorm(1)-pnorm(0),pnorm(2)-pnorm(1),pnorm(3)-pnorm(2),pnorm(4)-pnorm(3),1-pnorm(4))

res\_4 <- chisq.test(x = W.freq, p = W\_norm)

#Chi-squared test for given probabilities  
  
#data: W.freq  
#X-squared = 6.3506, df = 9, **p-value = 0.7044**  
#  
## Finding the limit  
plot(to\_test,W,type = "l",main = "W vs M values",xlab = "M values",ylab = "W values")

Figure4.3 - W values versus M values



# Conclusion

To conclude, I utilized the Random values on R and performed statistical analysis using R. I generated the different random values and used Chi-Square Goodness of Fit Test to verify whether the generated values belong to a particular probability distribution. Finally, I answered the below question accordingly :

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 Exponential Random Variables has the Gamma Probability Distribution
3. The output of the algorithm of problem 3 has a Standard 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 Standard 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 about 1.30 and Standard deviation is equal to 0.091. It seems that as the number of Random values increase, W tends to get close to 1.30.

# References

Relationships among probability distributions. (2019, December 7). Retrieved from

<https://en.wikipedia.org/wiki/Relationships_among_probability_distributions>