**Module 1: Week 1 Introduction to Enterprise Analytics**

**Author’s Name:** Shivani Sharma

**Course Number:** ALY6050

**Course Title:** Intro to Enterprise Analytics

**Academic Term:** Fall 2019 CPS Analytics

**Instructor’s Name:** Dr Alex Huang

**Assignment Completion Date:** 01-15-2020



**Introduction**

This assignment is based on the logic of generating random variables and applying the statistical analysis on those random values. After obtaining the results we have to plot relative frequency histograms, probability distribution, probability plot and have to perform chi-squared of best fit for each of the problem statement. Relative frequency histogram represents values in percentages. Probability distribution provides all possible values within a given range. Probability plot provides a difference between 2 data sets by comparing them.

Chi-squared test is a comparison of histogram with density and is applied to unidimensional distribution in order to calculate the cumulative distribution. Basically, in layman’s terms it is performed to decide if there is any difference between observed and expected values. Following is the formula used for calculating the chi-squared test statistics:

𝝌 𝟐 = ∑ (𝑶𝒊−𝑬𝒊 )^2/ 𝑬𝒊

Where Oi and Ei are the observed and expected frequencies and degree of freedom can be calculated using the formula df=K-p-1

Where k is the number of bins and p is the number of parameters which are estimated.

Chi squared test is right tailored test i.e following condition should always meet:

𝑷. 𝒗𝒂𝒍𝒖𝒆 = 𝑷(𝜲 𝟐 ≥ 𝝌 𝟐 )

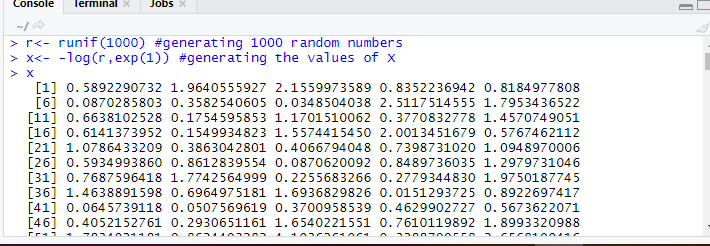
I would like to explain all the problem statements one by one in analysis section with the logic and output results of each problem.

**Analysis**

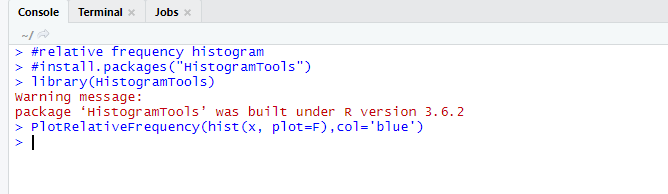
**Problem 1:** In this problem we have to generate 1000 random numbers and for each value we have to calculate x by using following expression:

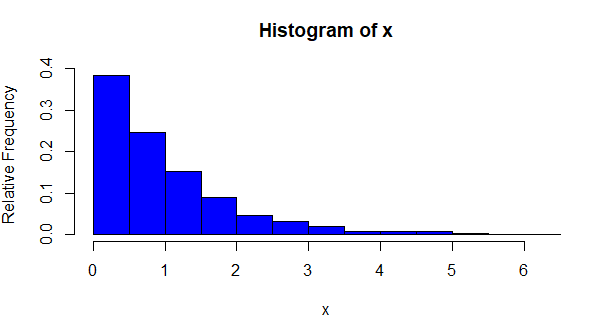
**𝒙 = −𝑳n(𝒓)**

Following is the logic and output for this function:



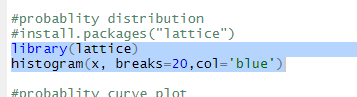
Now we have to plot a relative frequency histogram from the obtained values of “X”. For that I have used a inbuilt function in HistogramTools package

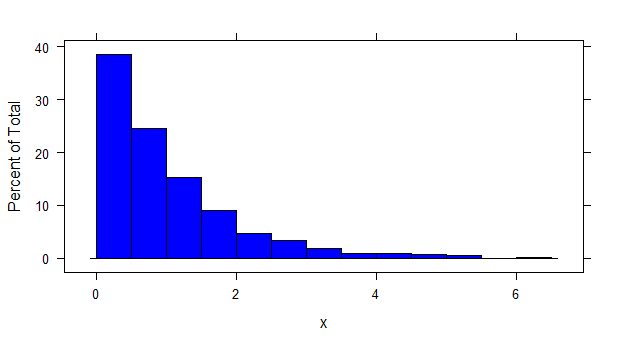




From the above histogram we can conclude that the relationship between x and relative frequency is negative and it is showing a shape of e^-x i.e exponential shape.

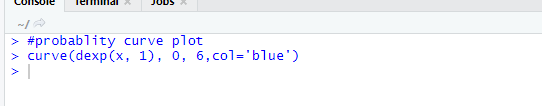
Now in support of the results obtained above in the relative frequency histogram we can verify the relationship by plotting the probability distribution for the parameter “X” and here is the logic for that:

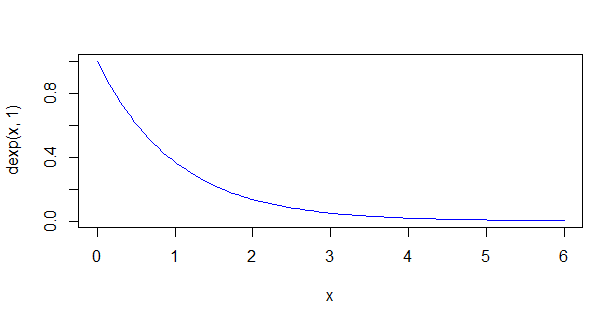




From the above graph we can conclude that the relationship between percent of Total and X is exponential i.e values are decreasing continuously exponentially which is best fit for X.

Now the above assertion and results can be supported by creating a probability plot curve for the X which is showing the same relationship i.e a exponential graph curve i.e probabilities are continuously decreasing with increase in value of x.



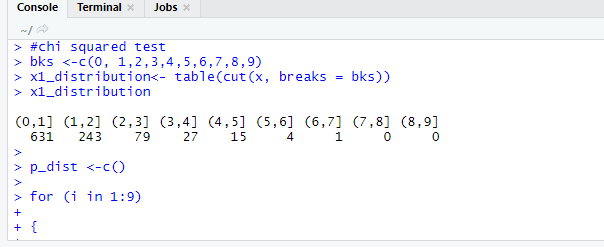


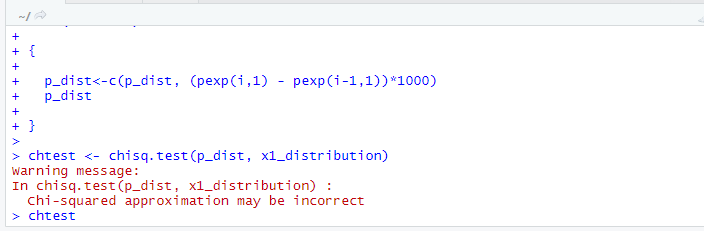
From the above graph we can conclude that the relationship is e^-x i. exponential one.

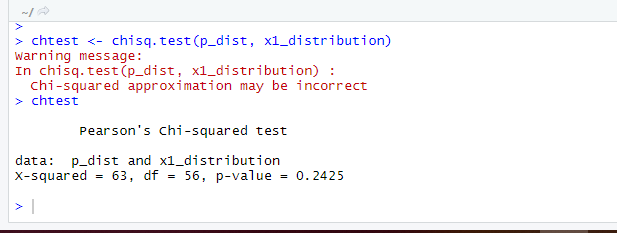
Now in further addition to the support of our conclusions drawn.

Let’s take Ho as a hypothesis that there exist a exponential relationship among the values.

Ha exponential relationship do not exist among values. Level of significance is provided as .05







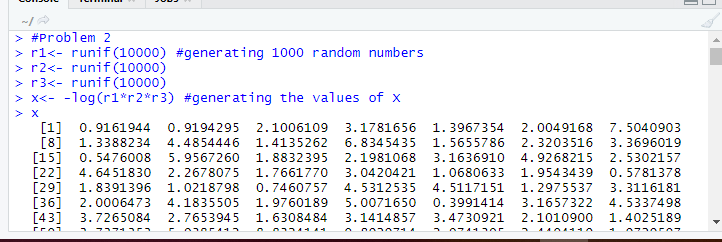
From the results of chi squared test we can observe that the p-value is 0.24 which is more than the significance level. So, we are accepting the null hypothesis and hence there exist a exponential relationship in the distribution.

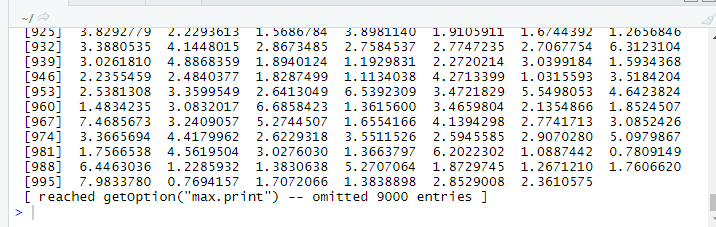
From this problem we have observed that the pattern or relationship for a function shows inverse relation for a negative logarithmic function.

**Problem 2:** Now coming to problem 2 in which we have to generate 3 random numbers with 10,000 random values each. And then calculate the value of following function using the values of random numbers generated:

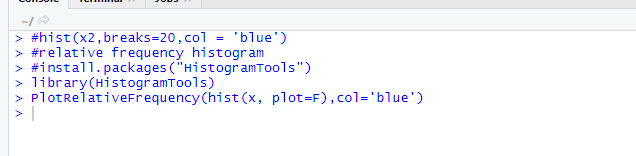
𝒙= −𝑳n(𝒓𝟏𝒓𝟐𝒓𝟑)

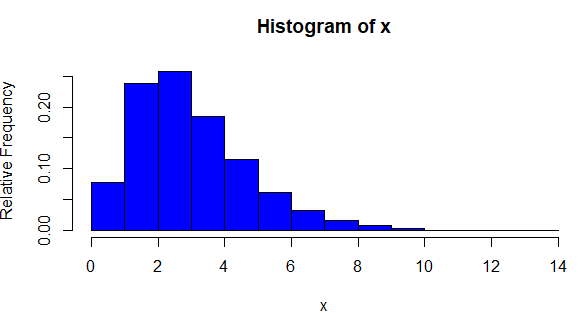
Following is the logic to calculate this value:



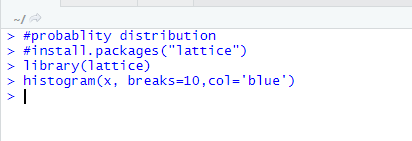


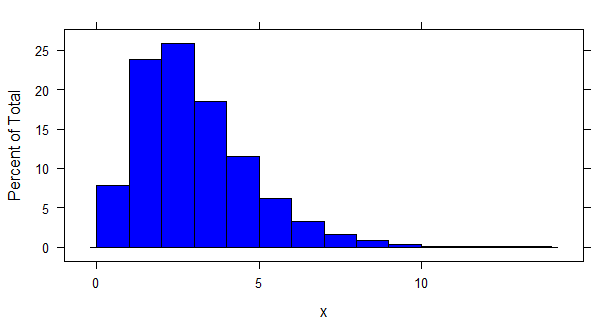
Now in order to explain the relationship between the values we can plot a relative frequency histogram for the function and following is the logic:





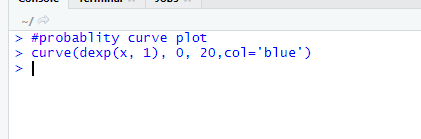
From the above histogram we can observe that there exist a right skewness in the bell shaped curve. In further addition to support this observation we need to plot a probability distribution

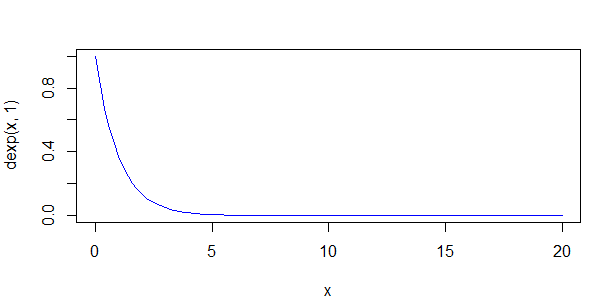




From the above plot we can conclude that there is also a right skewness in the bell shaped curve.

Now to further support this assertion we need to create a probability plot so as to make the assertion more stronger.





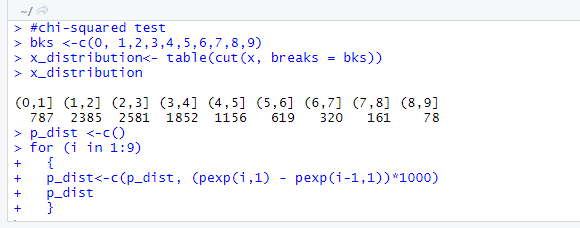
From the above curve we can clearly observe the right skewness too in the shape. So, the relationship is showing a pattern of exponential pattern.

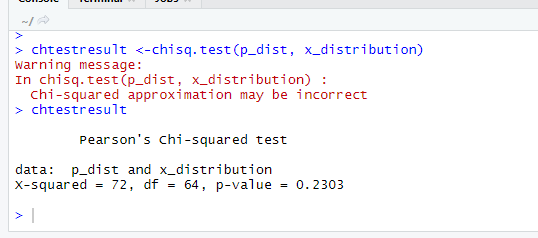
Now coming to further addition to assertion validation we will do a Chi-squared test for best fit.

Let’s take Ho as a hypothesis that there exist skewness in the pattern of values.

Ha there do not exist skewness in the pattern among values. Level of significance is provided as .05.

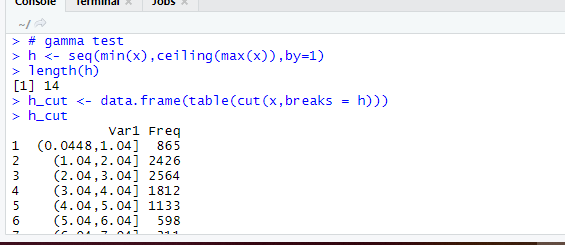
Here is the logic for the same:

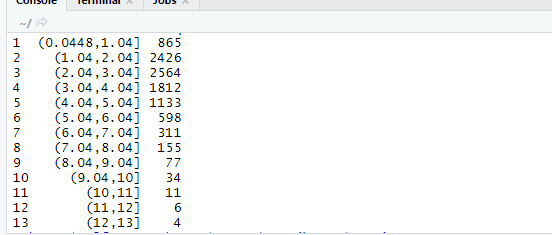


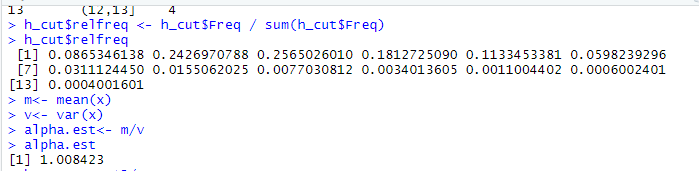


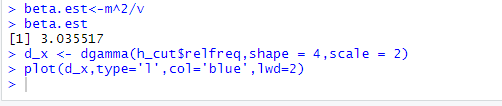
From the result of chi-squared test we can observe that the p-value is more than the significant value. So we are accepting the null hypothesis i.e the skewness exist in the values.

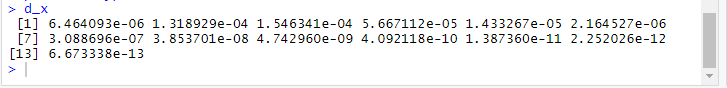
Now we have to perform a gamma test and plot the result. Following is the logic for the same:

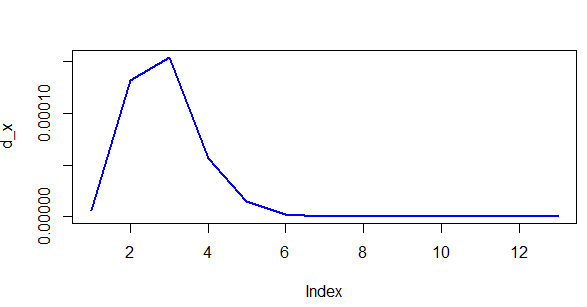












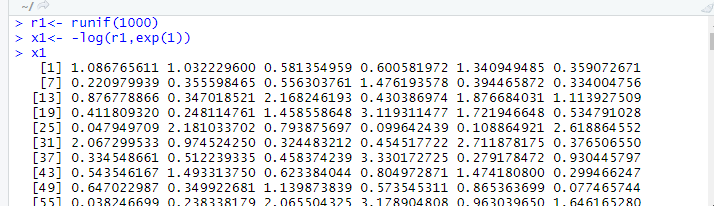
From the above results we can conclude that the result is showing a right skewness in the result obtained. The values are increased to a peak and maximum value which is showing a steep declination in the pattern and the results are hence right skewed.

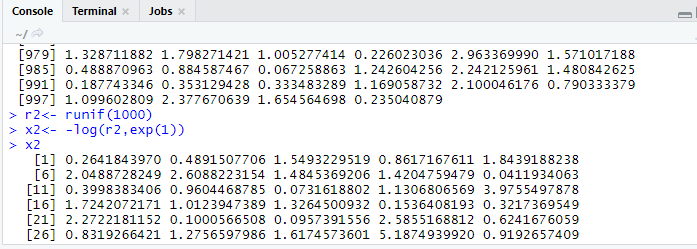
**Problem 3:** In this problem we have to generate a set of uniform random numbers with 1000 values and calculate negative logarithm for both separately as follows:

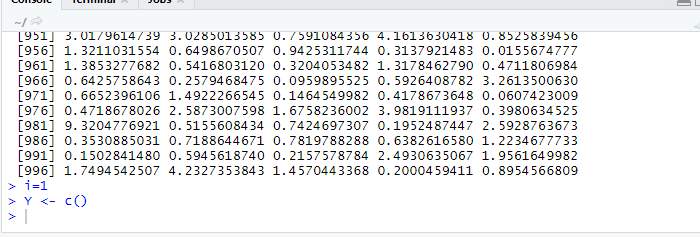
x𝟏 = −𝑳(𝒓𝟏)

x𝟐 = −𝑳(𝒓𝟐)

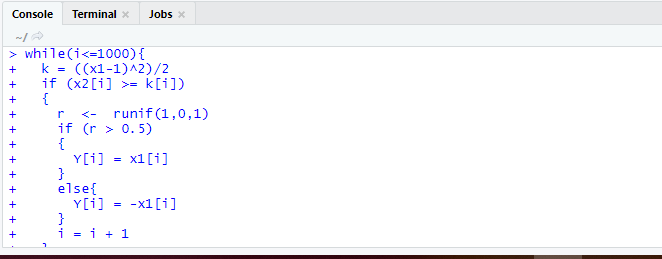
And then there is a condition provided we have to calculate a number K using the value of x1 using this formula: K=(x1-2)^2/2 and using this “K” value we will generate a random number only if x2 >=K and there are further conditions in the loop i.e if r>0.5 then we have to find the value for a new variable i.e Y which is equal to x1 otherwise we have to exit the loop. Now in order to achieve this logic I have used a “While” loop as I want to repeat the loop till the time condition is meeting and perform the calculation. Here is my logic:

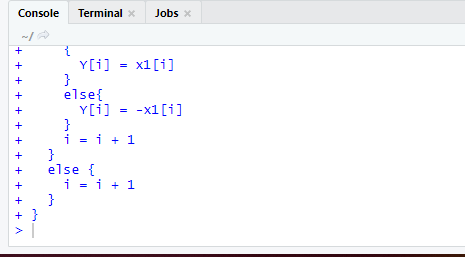




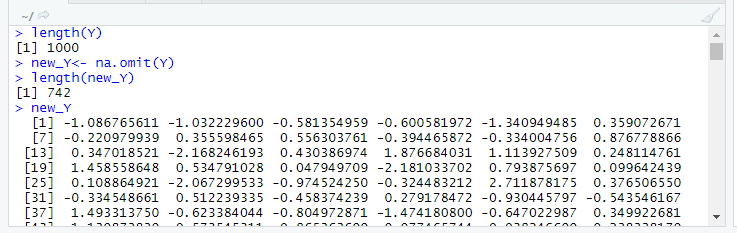


In the above logic I have just calculated the random numbers and initialized the new variable Y in which I want to store values after the condition is met as given in the problem statement.

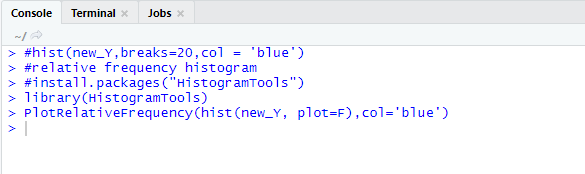




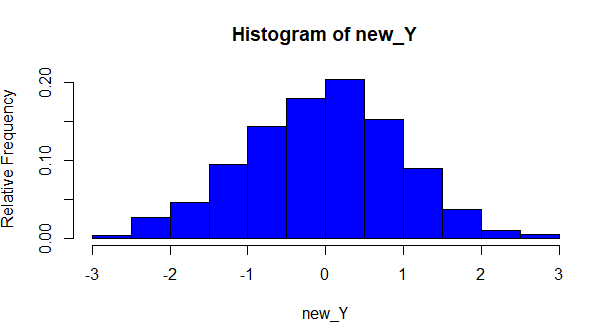
I have used cascading of if and else logic in order to meet the conditions and calculate the values. This loop will run every time and based on the values it will go into the next iteration.



Here are the values obtained for the Y value.

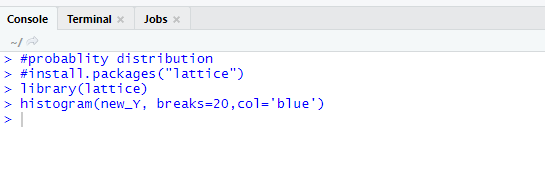


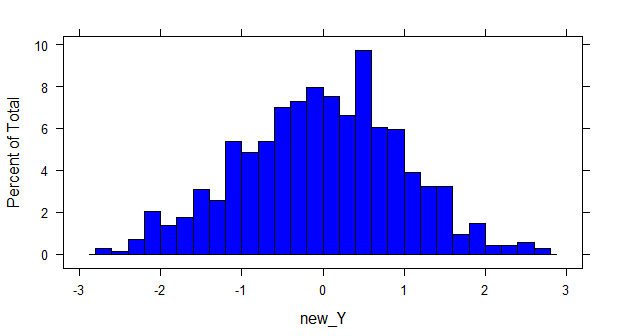
Now we will plot the relative frequency plot for the results obtained for the new value of Y which is as follows:



This shape is a bell curve shaped with slight left skewness in the pattern.

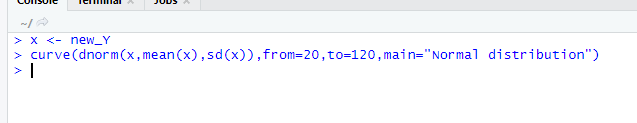
In further addition to this argument I have plotted a probability distribution for the same using a histogram which is as follows:

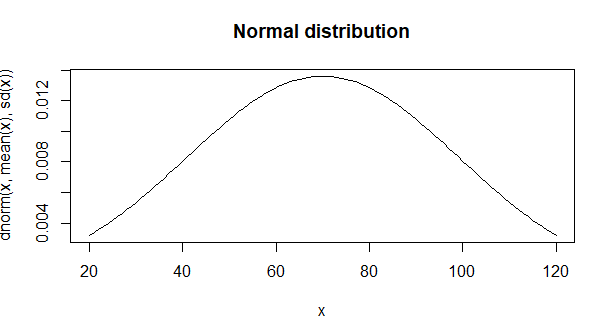




The relationship is again showing a of the pattern obtained.

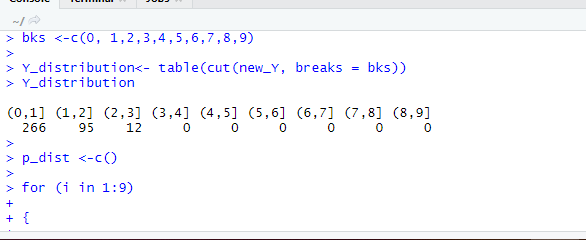
Now in further addition to this assertion we can have a one more plot i.e probability plot to represent the shape of the curve using dnorm function.

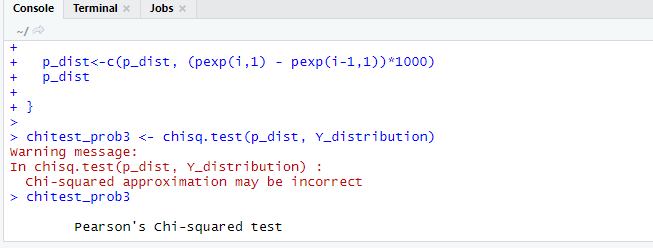


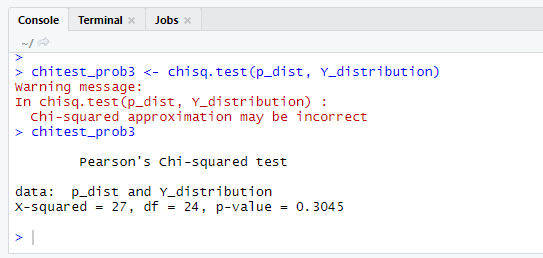


The probability plot is also showing a slight skewness, so our results and assertions are in sync with each other.

Now we have to perform the chi-squared test for the new value of Y obtained for finding the best fit. Let’s take Ho as a hypothesis that a left skewness is observed in the shape of pattern between values.Ha: there is no left skewness in the pattern among values. Level of significance is provided as .05.

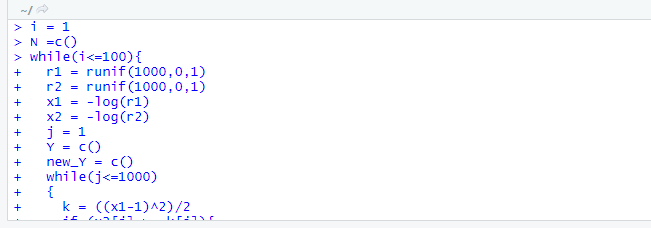


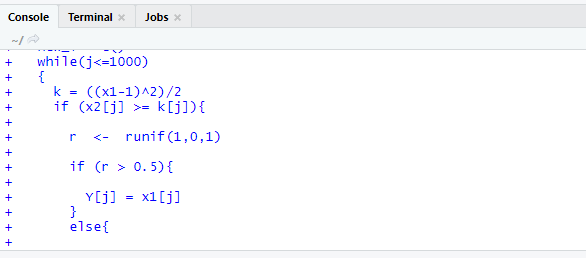


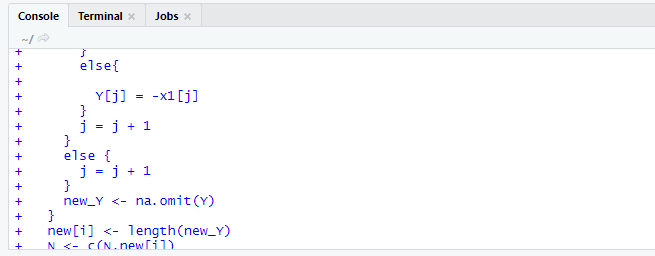


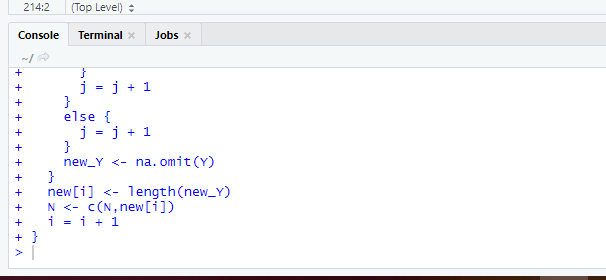
From the above results we can clearly see that the p-value is 0.3045 which is more than the significance level so we are accepting the hypothesis and hence our assertion was true i.e skewness exist in the pattern.

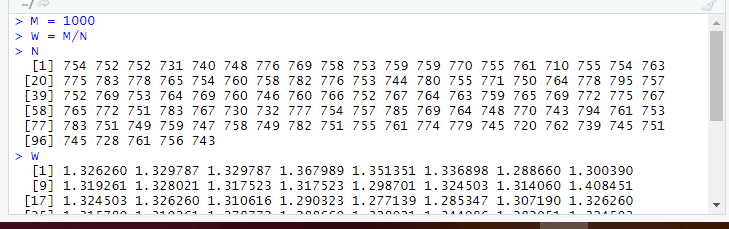
**Problem 4:** In this problemwe have to fill the gap of above problem i.e to consider and evaluate those random numbers of above problem which do not satisfy the condition x2>=K .So we have to create a new variable “W” and for which we have to find out the relationship between actual and expected values of W. Here is the logic to evaluate the condition and find out the value of “W”:

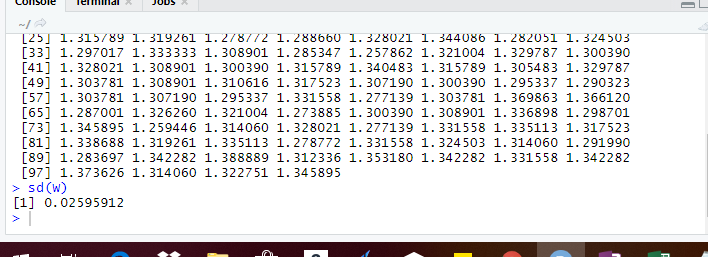






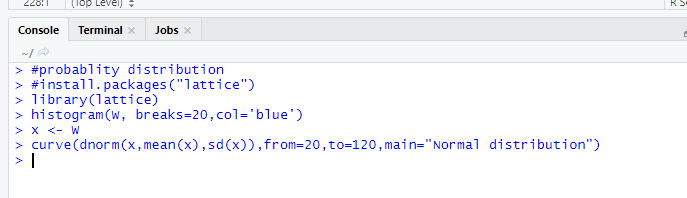


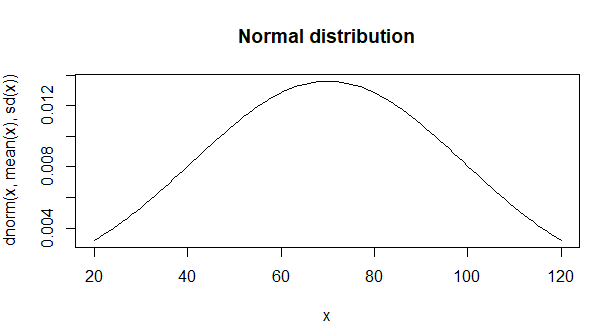




From the above logic we have obtained expected values of W and the standard deviation is coming out as 0.0259 for the values.

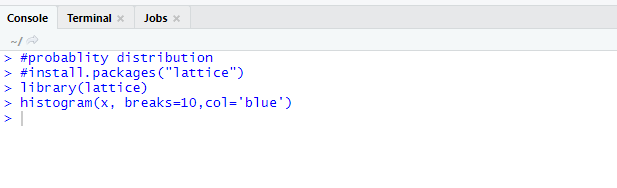
Now in further addition to this I calculated the probability distribution plot for the values obtained and the logic is this:

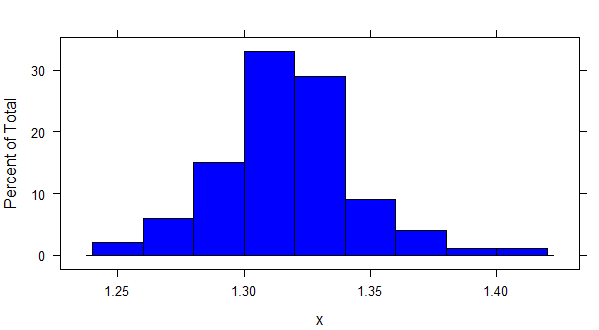




From the above results we can observe that the graph is a bell-shaped curve. i.e values of “W” are increasing constantly first and then there is a uniform decrease in the values.

Additionally, I wanted to have one more pointer to support this assertion so I decided to create a probability distribution histogram for the same which is as follows:



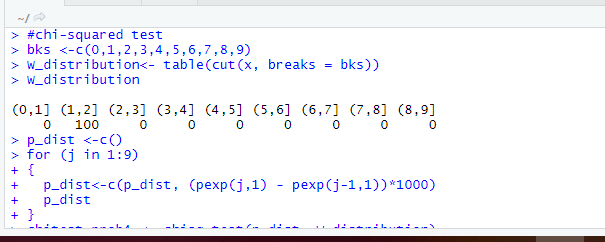


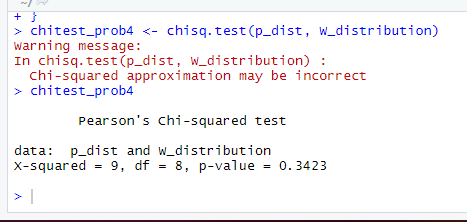
The pattern in the above graph will also be a bell-shaped curve if we connect the dots.

Now supporting the assertion further I did a chi-squared test for the “W”.

Let’s take Ho as a hypothesis that a uniform bell shaped curve is obtained for the values.

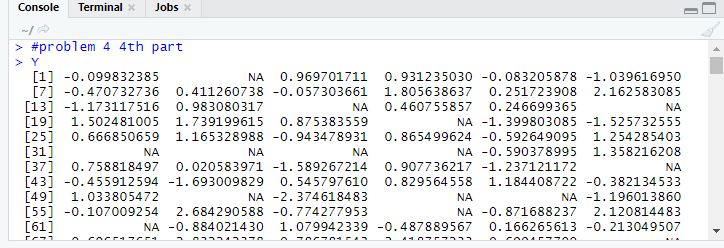
Ha: There is no uniform bell shaped curve is obtained for the values. Level of significance is provided as .05.

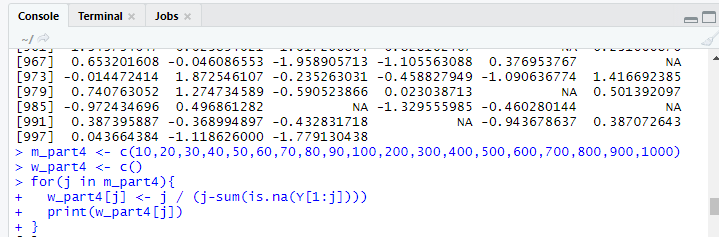




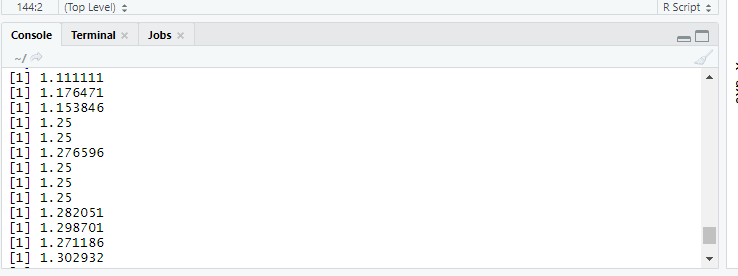
From the results of chi-squared test we can observe that the p-value is 0.34 which is more than the significant value i.e 0.05 so we are accepting the hypothesis i.e our results are true and the assertion is also strongly supported by the test.

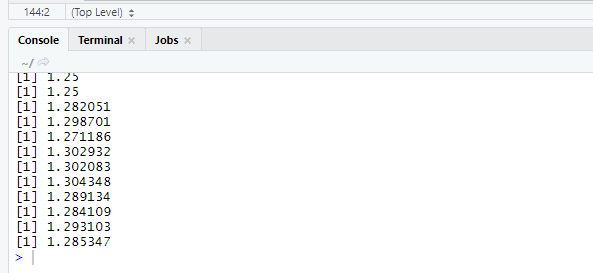
Now we are provided with certain values for M and we have to calculate the W values for them using the Y values we calculated earlier. So, here is the solution I used for calculating the values i.e storing it a vector and applying a simple “for loop” for all the values of “M”.





Here are the corresponding values of “W” for each value of “M”.





**Conclusion**

At the end I would like to conclude that the problems of the assignment gave a deep insight into probability concepts and helped to understand how to study the relationships in the trend observed for various distributions like probability distribution, relative frequency distribution. We observed different behaviors for all the problems i.e exponential, right skewed, left skewed and uniform bell-shaped too. Learned how to read the pattern and before concluding a final hypothesis test is required to decide the results.

Finally, the answers to the summary asked in the assignment:

1.If 𝒓 is a standard uniform random variable, then −𝑳(𝒓) has the **non-uniform exponential**

probability distribution.

2. The sum of three independent and identically distributed **uniform** random

variables has the **Right-skewness in** probability distribution.

3. The output of the algorithm of problem 3 has **a slight left skewed bell-shaped** probability

distribution.

4. In step 2 of the algorithm of problem 3, random variables 𝑿𝟏 and 𝑿𝟐 , each of whose

probability distribution is **normal**  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 **1.31237(mean of W)**.

# References

1. Stephanie. “Relative Frequency Histogram: Definition and How to Make One.” *Statistics How To*, 24 June 2018, [www.statisticshowto.datasciencecentral.com/relative-frequency-histogram-2/](http://www.statisticshowto.datasciencecentral.com/relative-frequency-histogram-2/).
2. “Stat Trek.” *Chi-Square Test of Independence*, stattrek.com/chi-square-test/independence.aspx.
3. “Using Chi-Square Statistic in Research.” *Statistics Solutions*, [www.statisticssolutions.com/using-chi-square-statistic-in-research/](http://www.statisticssolutions.com/using-chi-square-statistic-in-research/).
4. “Probability Plot.” *R*, math.furman.edu/~dcs/courses/math47/R/library/e1071/html/probplot.html.
5. <https://rstudio-pubs-static.s3.amazonaws.com/198393_3b9d87c451334863a0850be4ae605e39.html>
6. *QQ Plots and PP Plots*, homepage.divms.uiowa.edu/~luke/classes/STAT4580/qqpp.html.
7. Kabacoff, Robert. “Probability Plots.” *Quick-R: Probability Plots*, [www.statmethods.net/advgraphs/probability.html](http://www.statmethods.net/advgraphs/probability.html).
8. WillWill 75233 gold badges99 silver badges2121 bronze badges, et al. “How Do You Use Hist to Plot Relative Frequencies in R?” *Stack Overflow*, 1 Nov. 1960, stackoverflow.com/questions/4073223/how-do-you-use-hist-to-plot-relative-frequencies-in-r.
9. “Spatstat.” *Function | R Documentation*, [www.rdocumentation.org/packages/spatstat/versions/1.62-2/topics/CDF](http://www.rdocumentation.org/packages/spatstat/versions/1.62-2/topics/CDF).
10. “A Tutorial on Loops in R - Usage and Alternatives.” *DataCamp Community*, www.datacamp.com/community/tutorials/tutorial-on-loops-in-r?utm\_source=adwords\_ppc&utm\_campaignid=1658343521&utm\_adgroupid=63833879975&utm\_device=c&utm\_keyword=+loops +r&utm\_matchtype=b&utm\_network=g&utm\_adpostion=1t1&utm\_creative=319519154124&utm\_targetid=aud-299261629574:kwd-589281898934&utm\_loc\_interest\_ms=&utm\_loc\_physical\_ms=1018127&gclid=CjwKCAiA6vXwBRBKEiwAYE7iSxhVoK\_ARCokP-BFG-97N2ujuHGg13B5OSVkZFW3dFkgFEcB6wsMcBoCZ5kQAvD\_BwE.
11. “Skewness.” *Skewness | R Tutorial*, [www.r-tutor.com/elementary-statistics/numerical-measures/skewness](http://www.r-tutor.com/elementary-statistics/numerical-measures/skewness).
12. <http://www.r-tutor.com/elementary-statistics/probability-distributions/chi-squared-distribution>
13. <https://s3.us-east-1.amazonaws.com/blackboard.learn.xythos.prod/5a3148150d016/5145575?response-content-disposition=inline%3B%20filename%2A%3DUTF-8%27%27Examples%2520of%2520Probabilities%2520Calculations-R%25281%2529.pdf&response-content-type=application%2Fpdf&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Date=20200115T011851Z&X-Amz-SignedHeaders=host&X-Amz-Expires=21599&X-Amz-Credential=AKIAIL7WQYDOOHAZJGWQ%2F20200115%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Signature=2e8f0b0cfd5e12d8bd41adc4dc138e96068fa681817e05d8e0e8fbbdc9a69fca>