Chi-squared goodness of fit test

Rebecca Leu

Northeastern University

Introduction to Enterprise Analytics

Introduction

For this project we will be preforming statistical analysis and selecting probability distributions that fit the datasets we are looking at. To see if the probability distribution that was chosen for each dataset is a correct assumption, we will be preforming a goodness of fit test using the chi-squared test of best fit with a 0.05 level of significance. All the work is done in an excel spreadsheet ALY6050-Module 1 Project – Leu – R.

Analysis

Problem 1

After completing the random number generation R and generating X values from finding the negative natural log of each R value, a histogram is completed for the data. The histogram gives a great visual representation of the X values to see what type of probability distribution this follows. When looking at the histogram I compared the shape of the distribution with page 85 of the text showing the exponential distribution and concluded that this dataset followed the exponential distribution. Once I decided upon a distribution, I computed the expected frequency for each bin and that ran a chi-test using the CHISQ.TEST formula. The formula gave a significant p-value. This concludes that the X’s in this dataset follow the exponential distribution within a 0.05 level of significance.

Problem 2

Problem 2 created 3 lines of random variables and multiplied them by the negative natural log giving us our X data to work with. After completing a histogram of the data, it was easy to see the pattern in the data followed the gamma distribution. The data was skewed to the right with the highest bin having 124 frequency. I also completed a probability plot of X and found it followed that same gamma pattern. Following the same steps we took on problem 1, but replacing the exponential distribution function to the gamma distribution function I calculated the expected frequency of the data following the gamma distribution. I used 3 and 1 as the Alpha and Beta from the hints and theoretical background explained project directions. After completing both the frequency column and the expected frequency column, I again ran the CHISQ.TEST formula which gave a very low p-value confirming that the X’s in this problem followed the Gamma distribution.

Problem 3

In problem 3 we created an algorithm to work the random numbers though that eventually gave us our Y values to work with. The 2 pairs of random variables, 1000 each, were first generated into the same X values we used in problem one, which ended up being an exponential distribution. After our X calculations, we calculated k and generated another random dumber if k was greater than x2. If the random number was less than 0.5 we accepted x1 as y, if it was greater than .5 we accepted -x1 as y. but since there were sometimes instances that x2 was not greater than k we did not obtain any number. Because of this, after we went through the algorithm 1000 times, we only ended up with 777 Y values to work with in this problem.

After running the algorithm, we took the 777 Y values and followed the same process as in the previous problems. First creating a relative frequency histogram and selecting a probability distribution that followed the best fit for Y. It was determined by the shape of the histogram that this set of values followed the normal distribution. Again, the normal probability expected values were computed and a CHISQ.TEST was completed which resulted in a low p-value. The p-value from the chi best fit test confirmed our Y dataset followed that normal probability distribution.

Problem 4

Finally, problem 4 we looked again at the Y values used in problem 3 and went through a couple of additional steps to create our W dataset. The W dataset consisted of M, the number of iterations needed to generate the N of an accepted Y value. Creating two columns for M and N, the W column was a formula M/N for each line. My Y values ended up computing values for the first 12 number, so our W dataset had an interesting start, all equaling 1 for those firth 12 values. After that the values took an interesting turn which overall looked to be a rotated gamma distribution. The histogram showed low values at the beginning with a spike and decrease at the end. Overall, the values appeared to be approaching 1.32 but they never quite reached there. The table below shows how the values increased then leveled out over time. As with the other problems I computed the expected values based off the gamma distribution and computed the chi test p value which confirmed the idea that this dataset also followed the gamma probability distribution. Although this distribution was skewed to the left instead of the right, it still followed the same pattern as the normal gamma distribution.

|  |  |
| --- | --- |
| M | W |
| 10 | 1 |
| 20 | 1.12 |
| 30 | 1.12 |
| 40 | 1.18 |
| 50 | 1.23 |
| 60 | 1.23 |
| 70 | 1.26 |
| 80 | 1.25 |
| 90 | 1.24 |
| 100 | 1.27 |
| 200 | 1.30 |
| 300 | 1.30 |
| 400 | 1.28 |
| 500 | 1.29 |
| 600 | 1.28 |
| 700 | 1.29 |
| 800 | 1.28 |
| 900 | 1.28 |
| 1000 | 1.29 |

Conclusion

Overall, the project helped to understand how to calculate different data using random variables, as well as analysis of the data to find patterns and sorting those patterns into probability distributions. Practice working in Excel as well as preforming complex formulas helped to make me feel more confident in my excel skills. Ultimately, I gained a better understanding of the system, as well as understanding probability distributions in more depth and understanding how a chi-square test can solidify which distribution the values belong to.

1. If 𝒓𝒓 is a standard uniform random variable, then −𝑳𝑳𝑳𝑳(𝒓𝒓) has the exponential probability distribution.

2. The sum of three independent and identically distributed R 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 𝑿𝑿𝟏𝟏 and 𝑿𝑿𝟐𝟐 , each of whose probability distribution is exponential are used to generate a random value 𝒀𝒀 that has the normal probability distribution.

The random value 𝑾𝑾 that was discussed in problem 4, has the Gamma probability distribution. The expected value of 𝑾𝑾 is: 1.32.