**Lab 2**

The following exercises were performed using the instructor provided data set Lab2\_data\_sets.MTW. The data set consists of five samples from the following distributions: Normal, Exponential, Uniform, F(x) = x2, and one that is to be determined through the course of performing the following exercises.

These exercises are centered about the concept of constructing Empirical Cumulative Distribution Functions (ECDFs), and then visually comparing the ECDFs to more standard Cumulative Distribution Functions (CDFs) in an attempt to identify the parent distributions of the data.

An Empirical Cumulative Distribution Function is a step function that jumps by 1/n at every n data point. As an equation, it is represented by:

[1]

This function is used to approximate the true CDF of the data, and as more data is included, the estimator will converge to the true , by the law of large numbers.

1. Use the ECDF approach to find the Normal and exponential samples
   1. Sample 1



The Sample 1 ECDF does not appear to fit either distribution well; it is likely one of the three samples that do not come from a Normal or Exponential distribution.

* 1. Sample 2



Based on the plots, it would appear that Sample 2 came from a Normal distribution, the plot matches a normal CDF curve very well, but does not fit an Exponential CDF curve.

* 1. Sample 3



The Sample 3 ECDF matches the Exponential CDF curve well, but follows the Normal CDF curve poorly. This is likely the sample that came from the Exponential distribution.

* 1. Sample 4



The Sample 4 ECDF does not follow the Normal CDF well, and very poorly matches the Exponential CDF. This data likely came from one of the other three distributions.

* 1. Sample 5



Sample 5’s ECDF does not follow either CDF, Normal or Exponential, very well, and is a candidate for the product of one of the other three distributions.

1. Use the probability plot approach to find the Normal and Exponential samples (be sure your results are consistent with that of assignment 1)
   1. Sample 1



Again, this sample does not appear to come from either the Exponential or Normal Distribution, the plots are poor fits, being neither linear in trend, nor inside the lines of the 95% confidence interval. This result supports my assertion that this data does not belong to either Normal or Exponential distributions.

* 1. Sample 2



The Normal probability plot is a very good fit, with the data point trend being linear in shape and falling almost entirely within the 95% confidence interval. The Exponential plot is obviously a poor fit. These results add to my confidence in choosing Sample 2 as the product of the Normal Distribution.

* 1. Sample 3



Again, the Normal fit is very poor, but the Exponential plot displays a good match, again increasing my confidence in saying that Sample 3 came from an Exponential distribution.

* 1. Sample 4



Both of these probability plots exhibit poor fits, again leaving me to say that the data likely came from one of the other distributions.

* 1. Sample 5



Again, both plots exhibit poor fits with either Exponential or Normal distributions.

1. Use the quantile-quantile plot approach to find the Normal and exponential samples (be sure your results are consistent with that of assignments 1,2)
   1. Sample 1



The Exponential quantile-quantile plot does not display a good fit, but the Normal one indicates that Sample 1 might have come from the Normal distribution, but given the previous results I doubt that Sample 1 is Normally distributed.



When tested against the Uniform and F(x) = x2 distributions, it appears that the linear trend of points in the Uniform plot indicates that Sample 1 is likely Uniformly distributed. The data clearly did not come from the quadratic distribution.

* 1. Sample 2



The Normal quantile-quantile plot, does for the most part, display a linear trend in the points. The Exponential plot is not linearly distributed, so it is unlikely that the data is Exponentially distributed. These results further support the assertion that Sample 2 is Normally distributed.

* 1. Sample 3



The Normal quantile-quantile plot does not have a linear trend, so Sample 3 is clearly not Normal data. The Exponential plot only vaguely displays a linear trend, thus only weakly supporting the assertion that Sample 3 comes from an Exponential distribution.

* 1. Sample 4



Neither plot features the linear trend of points that indicates that the sample data may come from the distribution that we are testing it against, indicating once more that Sample 4 is neither Normally or Exponentially distributed.



When tested against the Uniform and F(x) = x2 distributions, it can be seen that Sample 4 matches the F(x) = x2 values well, producing the linear trend indicator that we look for in quantile-quantile plots, indicating that Sample 4 likely came from the quadratic distribution.

* 1. Sample 5



Both of these quantile-quantile plots do display a vaguely linear trend in the points, which makes it difficult to determine which distribution that Sample 5 might come from. This ambiguity does help to support the idea that Sample 5 comes from neither distribution, and likely comes from one of the three other distributions.



Sample 5 does not work well with either the Uniform or the F(x) = x2 distributions. It is likely Sample 5 is the data coming from the unknown distribution.

1. Find the theoretical 0.7 quantile of the exponential distribution with parameter 3; find the empirical 0.7 quantile of an exponential sample with the same parameter. Compare, explain and illustrate the difference in terms of the ECDF plot.

|  |  |
| --- | --- |
| 0.7 Quantile for an Exponential distribution with λ = 3 | |
| Theoretical | Empirical |
| 3.61192 | 3.58979 |



The Empirical quantile is only slightly less than the theoretical one. It is nearly undetectable in the graph above. This likely occurred because the sample data had a larger count of data than would be expected for an x value of 3.61 or less. This difference would likely decrease and be less likely to occur as the number of data included in the empirical sample increases.

This result helps to demonstrate that if an ECDF is created using enough data, it will closely match its parent distribution. This is in accordance with the property of the ECDF that as more data is used, the more the ECDF converges to the true CDF for that data.

1. Generate 100 RVs with CDF F(x) = 1-(1-x)3, x ∈ [0, 1]. Show the respective ECDF and the theoretical CDF in the same axes

