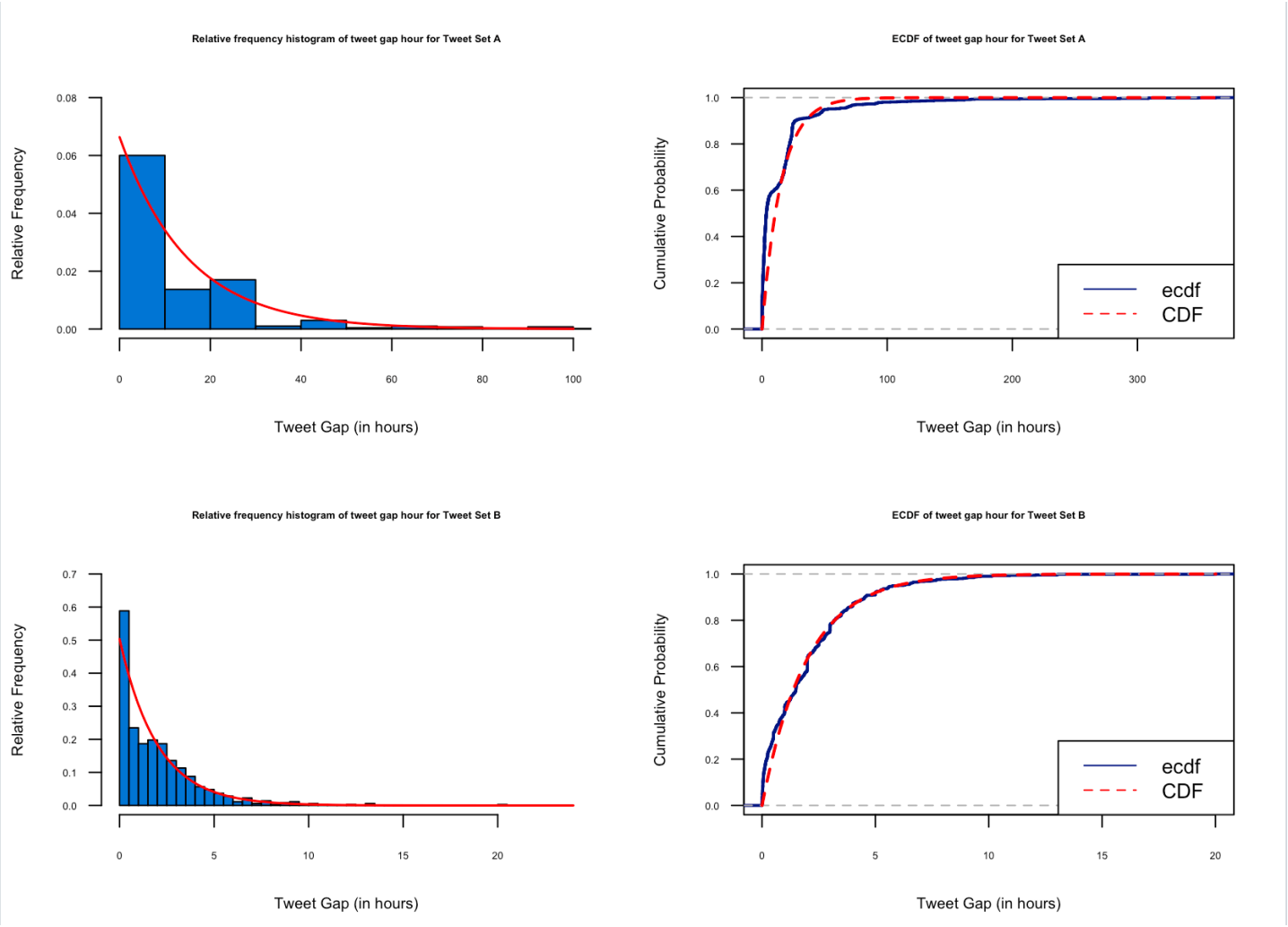
Analysis 3

I do not have concerns about measurement error in the first.tweet variate. This is because Twitter has a clear definition of the first tweet of the day. As the Twitter API creates the first.tweet variable we do not need to be worried about measurement error.

	Sample Size	Sample Mean	Sample Median	Sample Minimum	Sample Maximum	Sample SD
Tweet Set A	1168	15.0732	3.5014	0.0003	362.5881	31.0608
Tweet Set B	707	1.9884	1.4925	0.0003	20.0136	2.1846

The maximum value of tweet.gap.hour for Tweet Set B should not be greater than 24 because of the way the data is defined. We have defined Tweet Set B as "Just tweets that are not the first tweet of the day", this implies that the maximum value of tweet.gap.hour will calculate the time between the first tweet and the next tweet on the same day. As for every new day, we have to disregard the first tweet hence we do not go above the 24-hour time period. Hence the maximum value of tweet.gap.hour for Tweet Set B should not be greater than 24.



1(e)

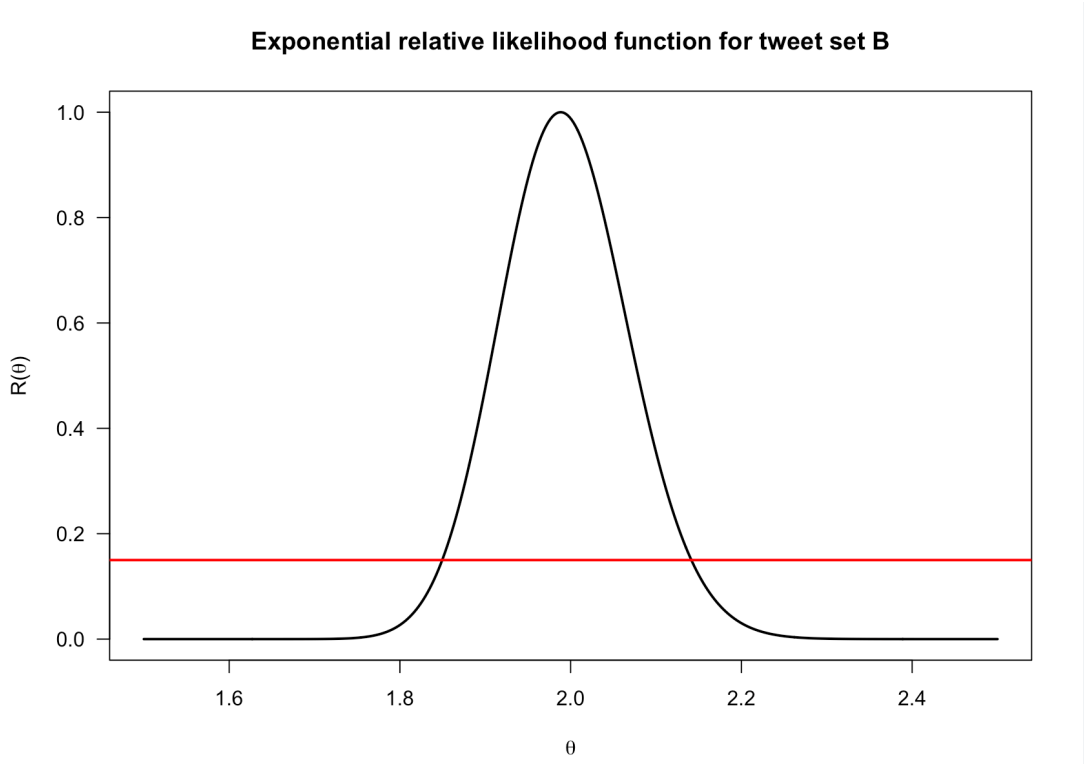
Tweet Set A: Based on the results in Analysis 1e, we can see that the relative frequency histogram has a right skew and the Empirical CDF of the data is similar to the data generated from an Exponential distribution. We would expect to see an Exponential distribution to have a similar pdf as our distribution's histogram, as evidenced by our graph in 1(e). As is evident by our graph in 1(e) the ECDF of our data and the CDF of the exponential function overlap in most parts but there are a few points where the ECDF goes above the CDF of the exponential function. Overall, the Exponential model fits our model well as seen by the graphs in 1(e).

Tweet Set B: Based on the results in Analysis 1e, we can see that the relative frequency histogram has a right skew and the Empirical CDF of the data is very

similar to the data generated from an Exponential distribution. We would expect to see an Exponential distribution to have a similar pdf as our distribution's histogram, as evidenced by our graph in 1(e). As is evident by our graph in 1(e) the ECDF of our data and the CDF of the exponential function overlap extremely. Overall, the Exponential model fits our model well as seen by the graphs in 1(e).

The maximum likelihood estimate of θ based on my sample is 1.988357.

Relative Likelihood Function Plot:



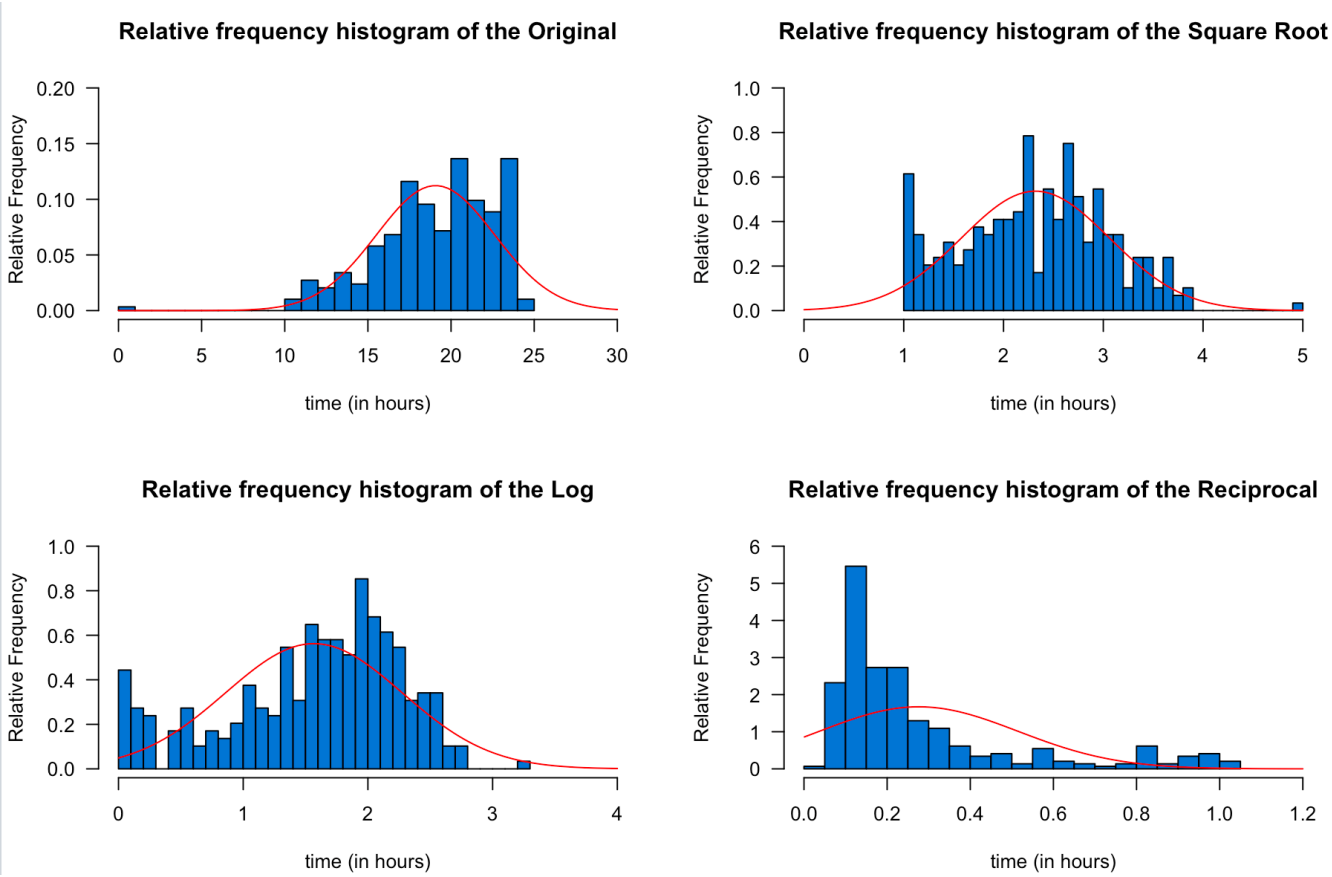
Caption

The 15% likelihood interval for θ is [1.8495, 2.1414].

The approximate 15%, 95% and 99% confidence intervals for θ are [1.9729, 2.0035], [1.8274, 2.1494], and [1.7768, 2.2000], respectively. These were calculated by calculating the c value (as we were using the CLT), where c = (1+p)/2*100% quantile of the normal distribution, here p is the confidence interval we need to calculate. We then find the upper and lower bounds using (Y' + c*sd/(n^1/2)) and (Y' - c*sd/(n^1/2)) respectively, where sd is the sample standard deviation, Y' is the sample mean and n is the sample size of the distribution.

The approximate 95% confidence interval is most similar to the 15% likelihood interval. This is what I would expect because of the likelihood ratio statistic, we can prove (as we have in class) that the 15% Likelihood Interval approximates the 95% Confidence interval.

The interval [1.8274, 2.1494] tells us that we are 95% confident that the value of the time gap between a randomly chosen tweet that is not the first tweet of the day and its previous tweet is between the interval of [1.8274, 2.1494].



2(b)

The Gaussian model appears to fit the Square Root transformation transformed data best, because of the three transformations the Square Root transformation fits the superimposed Gaussian probability density function curve the best.

I chose the Square Root transformation.

The sample size is 293, the sample mean is 2.319552, and the sample standard deviation is 0.7434937.

A 95% approximate confidence interval for μ is [2.234066, 2.405038]. This was calculated by the formula [Y' - b*S/(n^(1/2)), Y' + b*S/(n^(1/2))], where Y' is the mean of our sample, and S is the sample standard deviation, n is the length of the sample; b is equal to qt((p+1)/2, n-1), where p = 0.95.

This is a confidence interval because it is using a Gaussian model's pivotal quantity to find the confidence interval.

The interval [2.234066, 2.405038] tells us that we can say with 95% confidence that the mean of the square root transformation of any random sampling of time intervals between the first tweet of the day and the last tweet of the previous day will lie in [2.234066, 2.405038] interval.

A 95% confidence interval for σ is [0.6877734, 0.8091144]. This was calculated by the formula [{(n-1)*S*S/d}^1/2,{(n-1)*S*S/c}^1/2], where n is the size of the sample, S is the standard deviation and c is qchisq[(1-p)/2, n-1] and d = qchisq[(1+p)/2, n-1].

I would conclude that Alex's confidence interval will be wider than compared to the one I have calculated in Analysis 4f. This is because Alex has a smaller sample size which in turn results in more variability in their data.