**Assignment 3 LaTeX Template**

**Important**: In Analyses 1e and 3b the template provides space for a single image, as we recommend generating your 4 plots using `par(mfrow = c(2, 2))`. If you generate four separate plots these should all be included.

**Analysis 1**

**1a**: My ID number is 20892405.

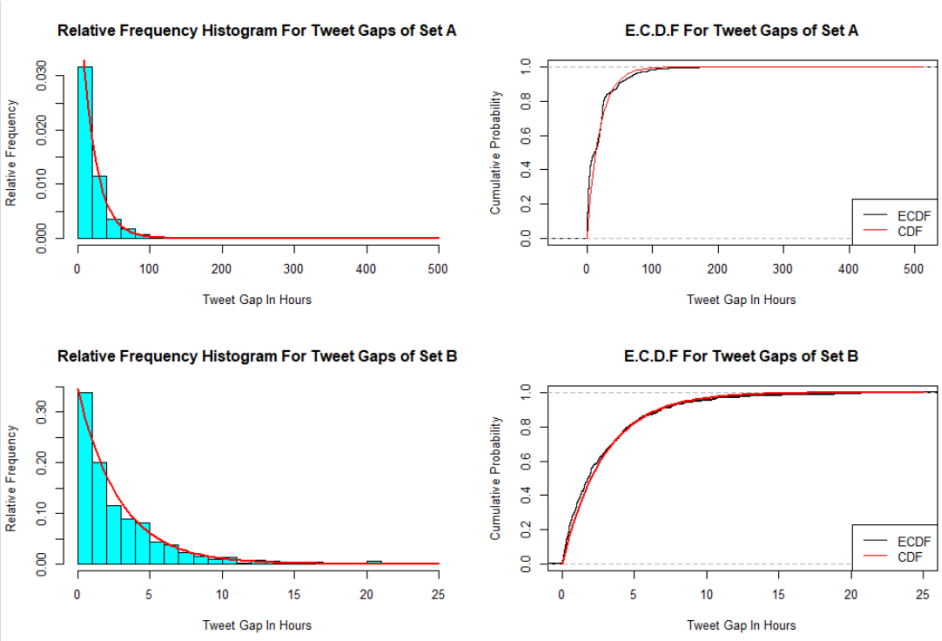
**1b**: I do not have concerns about measurement error in the **first.tweet**. This is because the first.tweet value for the corresponding tweet could have be taken from a web scrapper which is computer generated.

**1c**:

| | Sample Size | Sample Mean | Sample Median | Sample Minimum | Sample Maximum | Sample SD |
|---|---|---|---|---|---|---|
| Tweet Set A | 1133 | 19.363 | 11.287 | 0.0039 | 475.142 | 30.385 |
| Tweet Set B | 567 | 2.883 | 1.800 | 0.0039 | 20.791 | 3.315 |

**1d**: The maximum value of `tweet.gap.hour` for Tweet Set B should not be greater than 24 because for our Tweet Set B be we are removing the gaps between each day and only containing gaps within a single day.So it should make sense if the gaps are only accounted for within a 1 day time frame, the gap period should not exceed the hours of a day. .

**1e**:




**1f**: **Tweet Set A**: Based on the results in Analysis [1c/1e], we can see that the data seems to mimic an exponential model. There is a large collection of data points at the beginning and it seems to exponentially decrease as the get bigger. We can see a fairly large range as our range is approximately 435. We can see our data has a long right tail meaning the data is positively skewed which matches the characteristics of a exponential model. While for data generated from an Exponential distribution we would expect to see to have a similar fit with the data. The data points seem to line up with the super-imposed exponential curve with very little discrepancy. On our histogram we can see that

between 0 and 50 our data set is higher that the exponential model which is evident in our ECDF and CDF comparison. From our exponential model we would expect to see a lower values in gaps from the CDF. We can see that the ECDF and CDF are very similar in shape with little discrepancy, We can see that the median is very close in value. Overall, the Exponential model seems to fit out data well. .

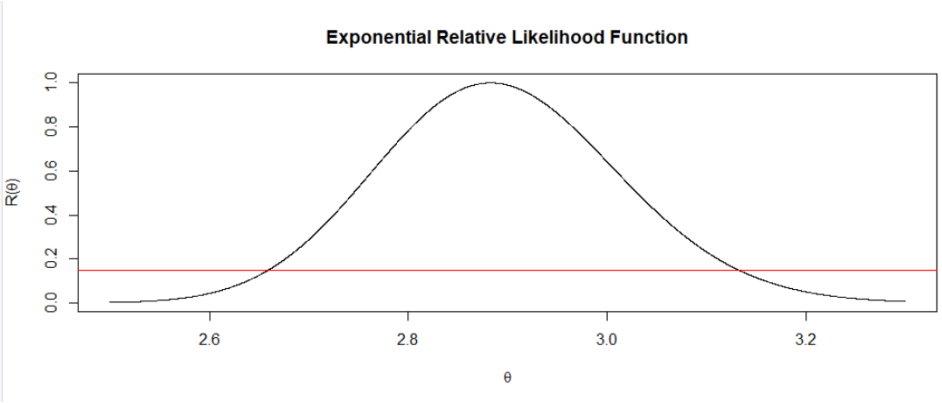
**Tweet Set B**: Based on the results in Analysis [1c/1e], we can see that the data seems to mimic an exponential model fairly well (Better than Tweets Set A). While for data generated from an Exponential distribution we would expect to see expect the same results as our data set with very little discrepancy. We would expect the data to be slightly lower but almost an insignificant amount. There also see to be now outliers for data points and not a large range of data points. We can see our data has a long right tail meaning the data is positively skewed which matches the characteristics of a exponential model. Our CDF and ECDF seem to match perfectly. Overall, the Exponential model seem to be an accurate fit for the data.
Overall, the Exponential model appears to be a better fit for Tweet Set B, because there are less discrepancies between our ECDF and CDF compared to dataset A. Also our median lines up between both curves and there is less uncertainty due to the of outlier data points.

**Analysis 2**

**2a**: My ID number is 20892405.

**2b**: The maximum likelihood estimate of $\theta$ based on my sample is 2.883.

**2c**: Relative Likelihood Function Plot:




**2d**: The 15% likelihood interval for $\theta$ is [2.6594, 3.1321].

**2e**: The approximate 15%, 95% and 99% confidence intervals for $\theta$ are 15% [2.859976, 2.905769], 95% [2.645581, 3.120164], and 99% [2.571019, 3.194726], respectively. These were calculated by where meanB and sdB are the mean and standard deviation of our Tweet Set B.

meanB + qnorm((1+0.95)/2)*(meanB/sqrt(sampleB))
2.905769 (UPPER BOUND)
meanB - qnorm((1+0.95)/2)*(meanB/sqrt(sampleB))
2.859976 (LOWER BOUND)
meanB + qnorm((1+0.99)/2)*(meanB/sqrt(sampleB))
3.194726 (UPPER BOUND)
meanB - qnorm((1+0.99)/2)*(meanB/sqrt(sampleB))
2.571019 (LOWER BOUND)
meanB - qnorm((1+0.15)/2)*(meanB/sqrt(sampleB))
2.859976 (LOWER BOUND)
meanB + qnorm((1+0.15)/2)*(meanB/sqrt(sampleB))
2.90576 (UPPER BOUND)

Based on our sample set we took our $\theta$ to be our MLE and found the standard deviation is also the MLE since its Exponentially distributed. We then found our confidence interval using the formula $\hat{\theta} \pm a\frac{\hat{\theta}}{\sqrt{n}}$ where a is our z score that $P(-a \leq Z \leq a) = p$ where p is our confidence interval.

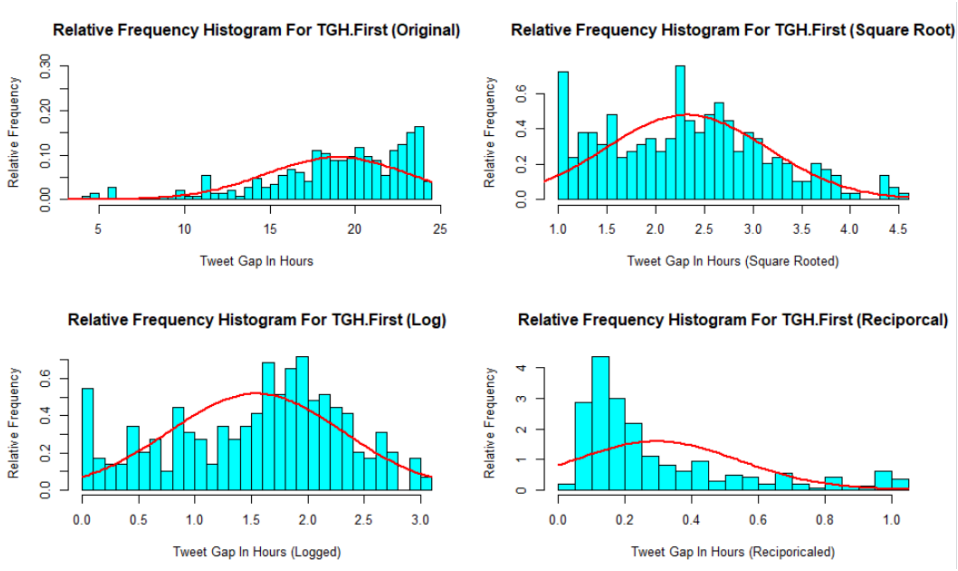
**2f**: The approximate 95% confidence interval is most similar to the 15% likelihood interval. This is what I would expect, according to Theorem 34 we would expect a 95% confidence interval form a 15% likelihood interval .

**2g**: The interval [2.61003, 3.155714] tells us that we are 95% confident that true average tweet gap in hours is between the interval 2.61003 hours and 3.155714. In other words we are 95% confident that the average tweet gap in hours is between [2.61003, 3.155714].

## Analysis 3

**3a**: My ID number is 20892405.

**3b**:




**3c**: The Gaussian model appears to fit the Square Root transformed data best, because there it is the graph with the least discrepancy between the relative frequency histogram and superimposed Gaussian curve. WE can see that the bars on the right hand size of the bell curve align with the superimposed curve (Between 2.5 and 4.0). We can see that the data is roughly evenly distributed with a slight long right tail.

**Analysis 4**

**4a**: My ID number is 20892405. In Analysis 3c I chose the Square Root transformation.

**4b**: The sample size is 292, the sample mean is 2.316283, the sample standard deviation is 0.8353614.

**4c**: A 95% [confidence interval/approximate confidence interval] for $\mu$ is [2.220069, 2.412498]. This was calculated by

meantf1 - qt((1+0.95)/2, length(tf1)-1)*(sdtf1/sqrt(length(tf1)))
2.220469 (LOWER BOUND)
meantf1 + qt((1+0.95)/2, length(tf1)-1)*(sdtf1/sqrt(length(tf1)))
2.412098 (UPPER BOUND)

Since our random variable is Gaussian we can find the confidence interval using $\bar{y} \pm b\frac{s}{\sqrt{n}}$ since $\mu$ and $\sigma$ is unknown

**4d**: This is an exact confidence interval, because we are using a Gaussian data .

**4e**: The interval [2.220069, 2.412498] tells us that we are 95% confident that the true average lies between [2.220069, 2.412498]. Note we are talking about the true average of the transformed data which is taking each gap and subtracting it from the max gap and adding 1. Then we take the square root of that value. Keep in mind the each gap is the gap between the first tweet of the day and the last tweet of the preceding day. .

**4f**: A 95% confidence interval for $\sigma$ is [0.7726569, 0.909228]. This was calculated by [explanation].

sqrt(((length(tf1)-1)*sdtf1)/qchisq((1-0.95)/2, length(tf1)-1))
0.9947996 (UPPER BOUND)
sqrt(((length(tf1)-1)*sdtf1)/qchisq((1+0.95)/2, length(tf1)-1))
0.8453752 (LOWER BOUND)
**4g**: I would conclude it would be narrower about the width of Alex's confidence intervals compared to those I calculated in Analysis 4f. This is because since we have more data points, then there is a lower uncertainty in our calculations. This also assuming our sample standard deviation and mean stay the same.