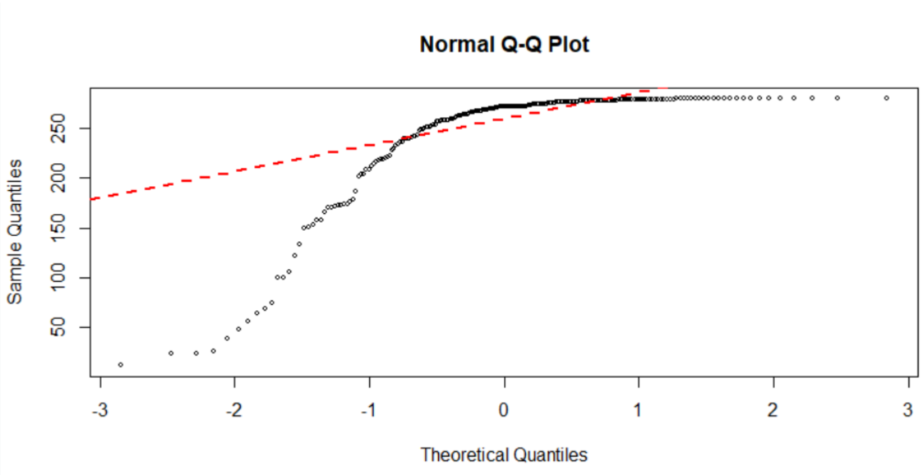
Assignment 2 LaTeX Template

Analysis 1

1a: My ID number is 20892405. I will be analyzing @theJagmeetSingh tweets.

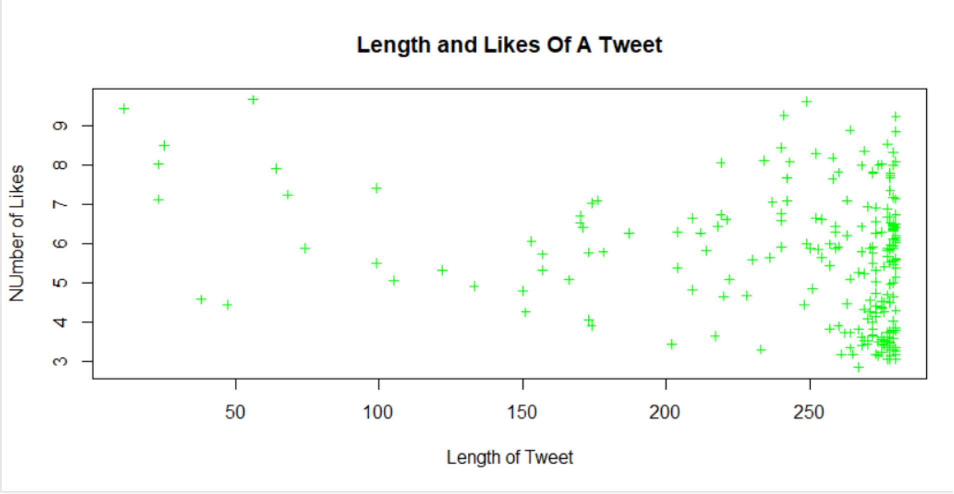
1b: I do have concerns about sample error in this study. This is because the type of account that is being used for our sample. A politician may tweet more on average than an organizational account and have a bigger following. There could be accounts who have a lower mean character count and higher likes mean. This can be compared to an account with a high character and likes mean. The activeness of the twitter account may affect the sample too. If an account tweets more frequently than another so they retain there audience.

1c: Q-Q plot of `length`:



1d: Based on the Q-Q plot we can see that the data does not fit a normal distribution as there is a big discrepancy between the line. We can see that the smaller length tweets are much smaller than expected when normal distrubed. This also shows us that our data is left skewed (negatively skewed) as we have a long left tail. There are alot of outlier data points. Most of the data is consist of higher length tweets. We can also see that the sample median and mean is higher than the line

1e: Scatterplot of `length` and `likes.log`:



1f: One way in which a scatterplot is not an appropriate summary for these data is that there is no correlation between the length and likes of a tweet. This graphical summary does not give a clear relationship between the 2 variates. We can see that the majority of his tweets are long but there is a variability in likes for his long tweets.

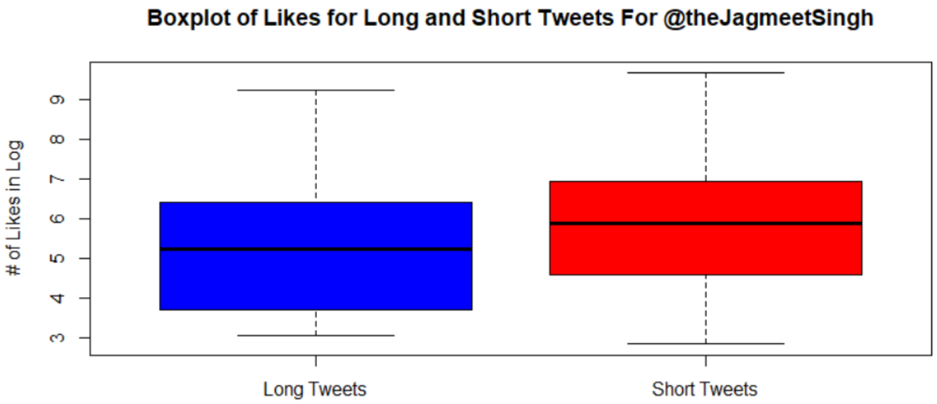
1g: The sample correlation is -0.2449277. This suggests that since the sample correlation is close to 0, there is no linear relationship between the 2 variates.

1h: The assumptions of a Binomial model are that each trial is independent and consists of 2 outcomes. For this variate, each tweet is independent from eachother and can either be a long or short tweet which is 2 outcomes..

1i: In the context of this study, θ represents probability of the tweet being a long tweet.

1j: The maximum likelihood estimate is 0.516. This was calculated by taking the sample size n = 225 and finding the proportion where the tweets are long tweets. Our MLE for binomial is also equal to $\frac{y}{n}$. That is $\frac{116}{225} = 0.516$.

1k: Side-by-side boxplot of `length.long` and `likes.log`:



1l: Based on the results of Analysis 1k, we conclude that short tweets receive more likes on average than long tweets. This is because see that the median is higher for short tweets and the interquartile range is higher compared to the long tweet.

Analysis 2

2a: My ID number is 20892405. I will be analyzing @theJagmeetSingh tweets.

2b: I do not have concerns about measurement error in the `media` variate. This is because the media value for the corresponding tweet could have be taken from a web scrapper which is computer generated.

2c: Summary statistics:

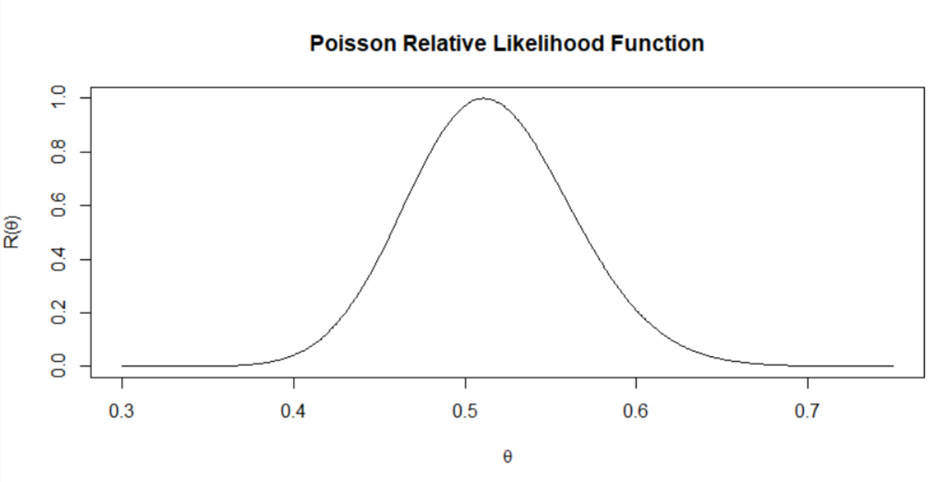
- Sample mean: 0.511
- Sample median: 0
- Sample mode: 0
- Sample standard deviation: 0.955s

2d: In the context of this study, λ represents the average number of media items in a randomly chosen tweet.

2e: The maximum likelihood estimate is 0.511. This was calculated by taking the mean of the sample media items of tweets.

2f: The maximum likeli hood estimate is 0.600. This was calculated by using dpois(0,$\hat{\theta}$) where the $\hat{\theta}$ is the MLE from 2e). We calculated this by using the invariance property by setting our $\lambda = MLE$ and solving $P(Y = 0|\theta = \hat{\theta})$ = dpois(0,$\hat{\theta}$)

2g: Relative likelihood function plot:

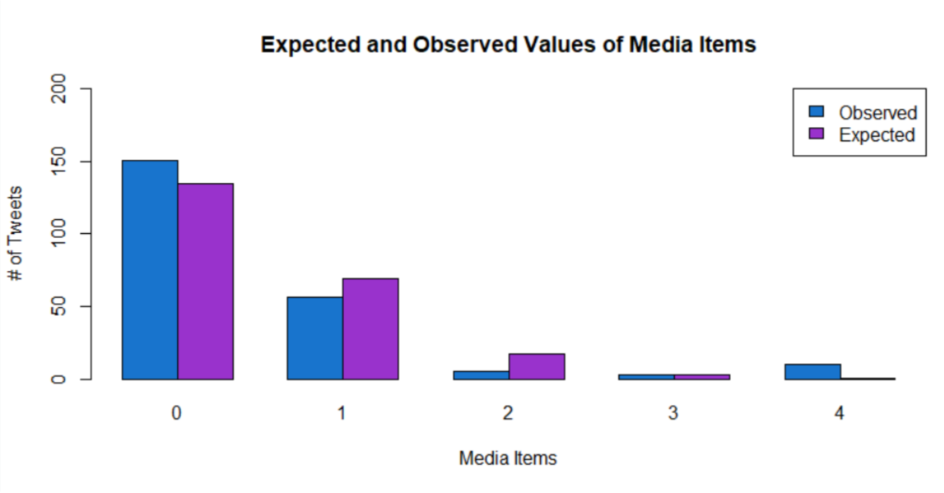


2h: $R(4) = 6.862283 \times 10^{-239}$. Based on this, we can say that it is a very implausible value as the value is very close to 0 making our observed data are very unlikely with $\theta = 4$.

2i: Summary of observed and expected counts:

media	Observed	Expected
0	151	134.961
1	56	68.980
2	5	17.628
3	3	3.003
4	10	0.384

2j: Grouped barplot of observed and expected counts:



2k: Based on the results of Analyses 2i and 2j, the Poisson model appears to fit the data as we can see the values between the expected values and observed values are close with little variability. We can also visually see on the bar plot, the similarity in shape and height between the observed and expected. In addition, Poisson is best modeled with discrete data that involves counting items and we are counting the number of tweets with a media item. Our media item take a discrete value.

2l: Summary statistics for `likes.log` and use of media:

Sample	Media use	
statistic	No media	Some media
Mean	5.525	5.576
Median	5.572	5.781
SD	1.785	1.385

2m: Based on the results of Analysis 2l, we conclude that having a media item has little no effect on the outcome of likes for a tweet. This is because we can see in the data there is a lack of variability between value.