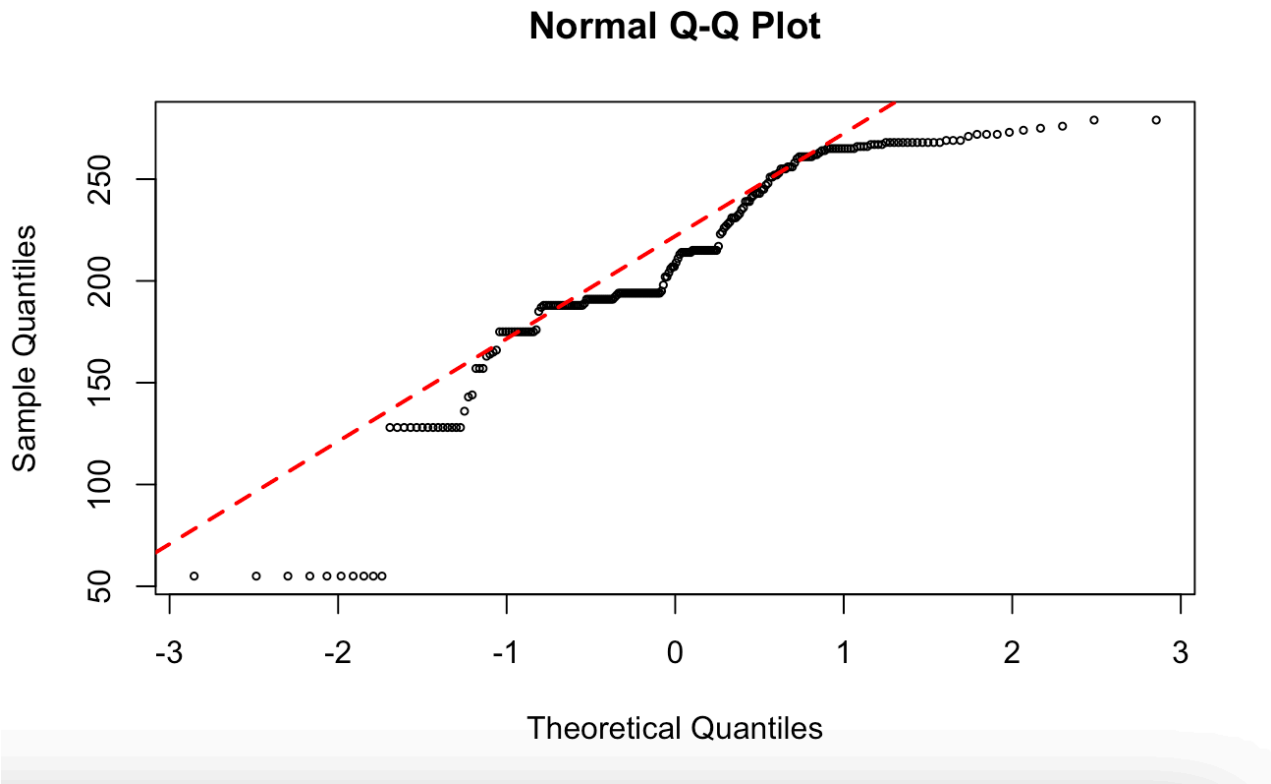
Analysis 2

I will be analyzing @CPHO_Canada's tweets.

I do have concerns about sample error in this study. This is because the Chief Public Health Officer of Canada is responsible for communicating a lot of complex public health decisions and facts with the common people hence they will require more visual aid as well as more explanation (thereby having more words). Thereby making their tweets longer.

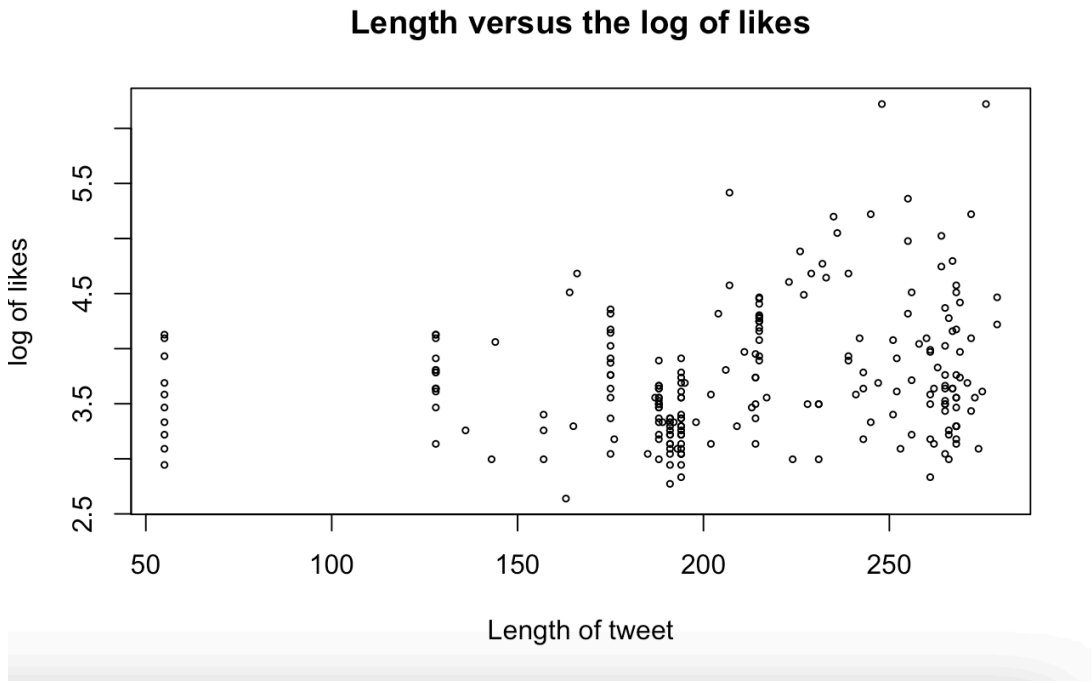
Q-Q plot of length:



Caption

Based on the Q-Q plot, we see that it has a staircase pattern, which is common for discrete data such as these. A Gaussian model does not seem appropriate for these data. We also see that there is a downwards U shape in the QQPlot, hence it has a Left skew. As the extreme values are very far from the theoretical Qualities we would imply that the graph has very heavy tails, hence the distribution has a kurtosis of greater than 3.

Scatterplot of length and likes.log:



Caption

One way in which a scatterplot is an appropriate summary for these data is that it can show how the relationship between the length go the tweets and the log value of likes. That is it can show if the data is positively correlated or negatively correlated. One way in which a scatterplot is not an appropriate summary for these data is if the linear relation between the two items is not strong then the graph does not a good representation of data.

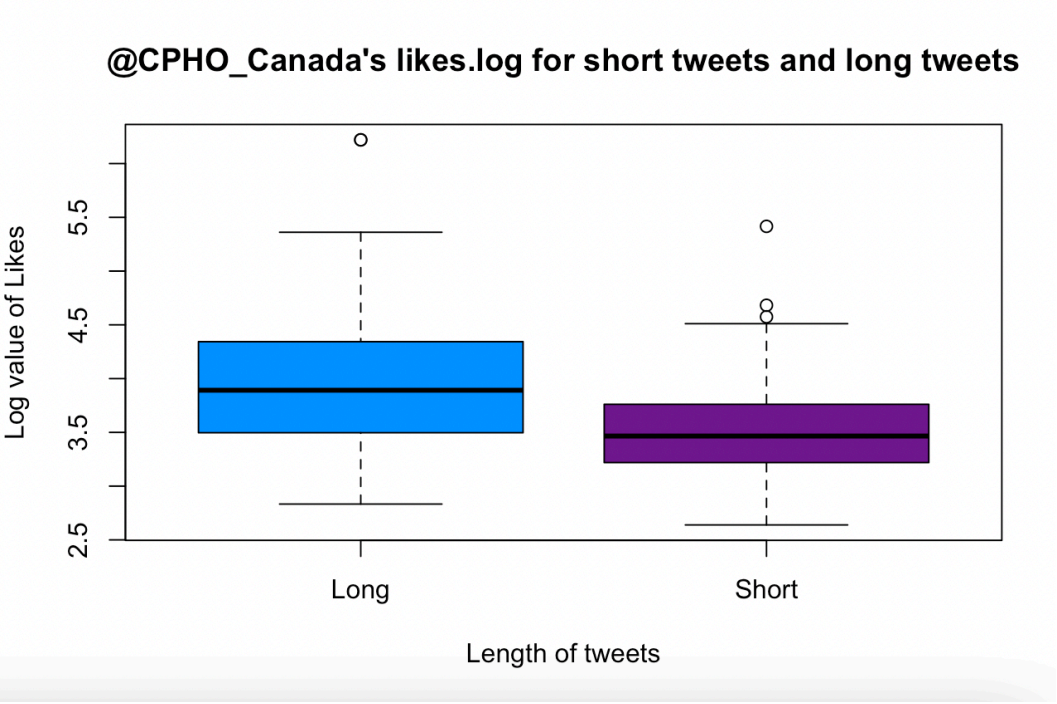
The sample correlation is 0.2207756. This suggests the relation between the tweets' length and the tweet's log of likes is slightly positive. This might suggest there is very slight evidence that as the tweets' length increases, the tweet's log of likes also increases.

A Binomial model assumes that we have independence in repeated trials with only two possible outcomes. For this variate, we know that if one tweet is long then that does not affect the other tweets and we are only allowed two possible options that are long and short.

In the context of this study, θ represents the probability of getting a long tweet in a random sample of n tweets.

The maximum likelihood estimate is [number]. This was calculated by [brief explanation].

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



Caption

Based on the results of Analysis 1k, we conclude that longer tweets receive more likes and shorter ones don't as is seen by the difference in the medians of the two box plots. This is because as the Chief Health Official of Canada, the user is expected to give a definitive explanation to the questions of the public. Hence the user writes longer tweets with a lot of explanation and they are equally liked by the public.

I will be analyzing @CPHO_Canada's tweets.

I do not have concerns about measurement errors in the media variate. This is because Twitter has clear distinctions on what it considers media items and a user has to import a picture/video [media item] to post it. Hence measuring error is not a problem in our data.

Summary statistics:

Sample mean: 0.7543103

Sample median: 1

Sample mode: 1

Sample standard deviation: 0.4314262

In the context of this study, λ represents the mean number of media items chosen in a random tweet.

The maximum likelihood estimate is 0.7543103. This was calculated by the formula that the maximum likelihood estimate for the Poisson distribution equals the mean of the distribution.

The maximum likelihood estimate of the probability a randomly chosen tweet contains no media items is 0.4703349. This was calculated by the following R code:

"thetahat <- mean(user$media)

theta <- seq(0, 1.5, 0.001)

table(user$media)

n <- dim(user)[1]

dpois(0, thetahat)"

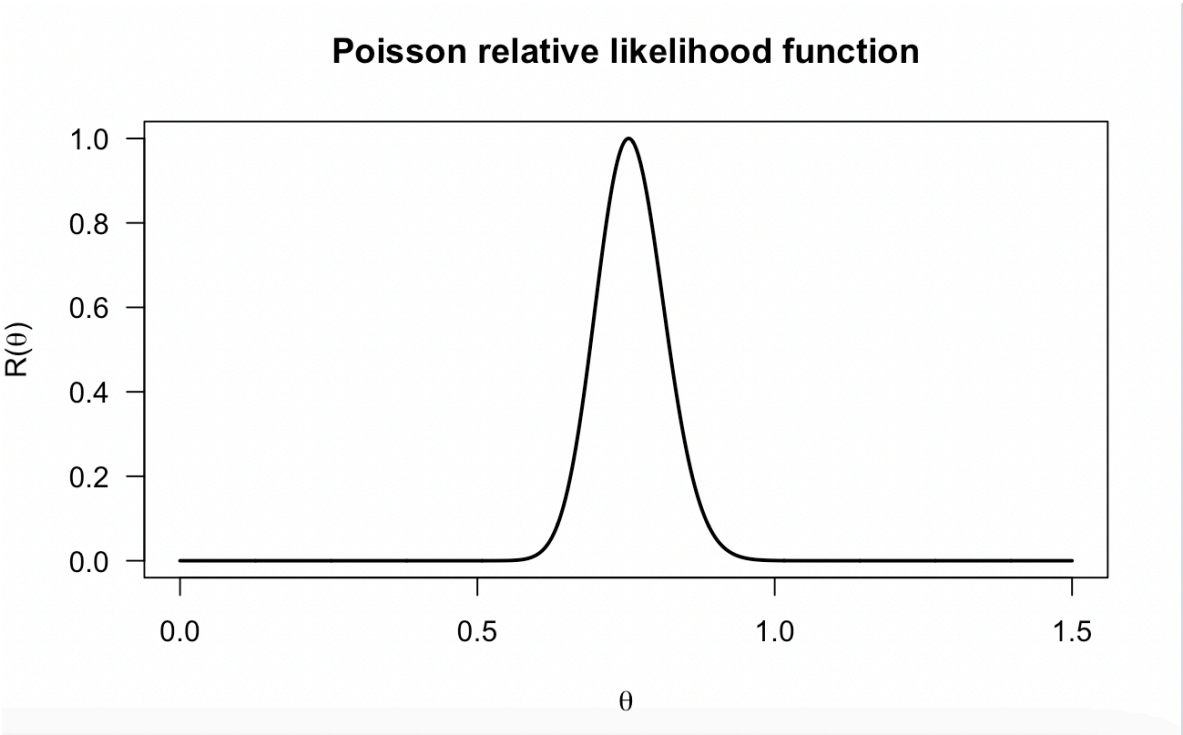
We see that by the Invariance property we can use the maximum likelihood estimate in the dpois function of 0 to find the maximum likelihood estimate of the probability a randomly chosen tweet contains no media items.

Relative likelihood function plot:
$R(4)$ = 5.827636*10^(-201). Based on this, we can say that the value of 4 is completely implausible for the value of lambda.

Summary of observed and expected counts:

media	Observed	Expected
0	57	109.118

1	175	82.309
2	0	31.043
3	0	7.805
4	0	1.472



Caption

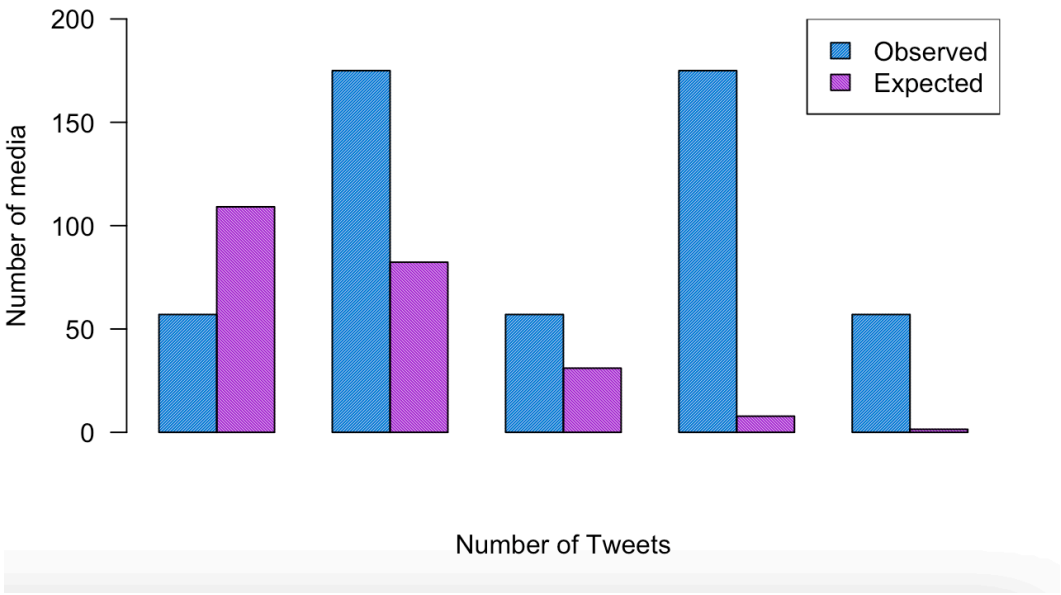
Grouped barplot of observed and expected counts:

Based on the results of Analyses 2i and 2j, the Poisson model appears to not fit the model. Hence, it is not an appropriate model for the media as the expected and observed values are very different from one another. In addition, the Poisson model works on the assumptions of Uniformity, Individuality and Independence. Uniformity is the presence of a constant probability of media in tweets, but the users are not required to use media, nor will they constantly use media. Hence Uniformity is not required by the model.

Summary statistics for likes.log and use of media:

Sample	Media use	
statistic	No media	Some media

Mean	4.14854	3.591287
Median	4.158883	3.496508
SD	0.6488849	0.5009712



Caption

Based on the results of Analysis 2l, we conclude that the tweets that contain do not contain any form of media are slightly more liked when compared to the tweets that have some form of media in them. This is because the median of no media is higher than the median of using some form of media. This is not always true and depends on the sample we have taken. I have mentioned how sampling error affects this user in question 1(b).