Problem

The target population for the study consists of all tweets in Canada (10 provinces and 3 territories) which will occur in the time period 1 January 2022 to 31 March 2022.

The study will analyze the following variates: *urls.binary, username, time.of.day, likes, retweets, long.words, hashtags.binary, media.binary, is.retweet.* The variate *urls.binary* is a binary variate, an attribute of interest for this variate is the proportion of tweets in the target population which contains at least one url. The variate *username* is a categorical variate, an attribute of interest for this variate is the average number of retweets in the target population for each username. The variate *time.of.day* is a continuous variate, an attribute of interest for this variate is the proportion of tweets in the target population which was tweeted between 1:00 a.m. and 2:00 a.m.. The variate *likes* is a discrete variate, an attribute of interest is the average number of likes for tweets in the target population. The variate *retweets* is a discrete variate, an attribute of interest is the mean number of retweets for tweets in the target population. The variate *long.words* is a discrete variate; an attribute of interest is the mean number of long words of tweets in the target population. The variate *hashtags.binary* is a binary variate, an attribute of interest for this variate is the proportion of tweets in the target population which contains at least one hashtag. The variate *media.binary* is a binary variate, an attribute of interest for this variate is the proportion of tweets in the target population which contains at least one media item. The variate *is.retweet* is a binary variate, an attribute of interest for this variate is the proportion of tweets in the target population which is retweeted.

One motivation question for analysis (1) is:

In the target population what proportion of tweets contain at least one url? This is a descriptive problem.

Two motivation questions for analysis (2) are:

Does the Gaussian model fit the variate time.of.day for @ONThealth in the target population? This is a descriptive problem.

In the target population, is the meam time of sending tweets in a day for @ONThealth the same as @GoAhealth? This is a descriptive problem.

Two motivation questions for analysis (3) are:

Suppose a tweet with 30 retweets is drawn from the target population. What is the approximate number of likes it will get? This is a predictive problem.

In the target population, if the logged number of retweets is increased by one unit, how will the number of likes changes? This is a descriptive problem.

One motivation question for analysis (4) is:

Suppose a tweet is randomly chosen from the target population, what is the probability that this tweet contains at least 2 long words? This is a predictive problem.

One motivation question for analysis (5) is:

Are the tweets in the target population that contain hashtags more likely to contain or do not contain media item? This is a descriptive problem

One motivation question for analysis (6) is:

In the target population, do the 8 different accounts have the same proportion of tweets that they sent are retweets from other accounts? This is a descriptive problem.

These data may not be used to examine any causative problems because the study is an observational study. An observational study in which the experimenter is not in control of the explanatory variates cannot usually be used to investigate a causative problem.

Plan

A suitable study population for this study would be all tweets by 8 Canadian provincial government's health agencies' social media teams(Alberta (@GoAHealth), British Columbia (@PHSAofBC), Newfoundland and Labrador (@HCS GovNL), Nova Scotia (@HealthNS), Ontario (@ONThealth), Prince Edward Island (@Health PEI), Quebec (@sante qc), and Saskatchewan (@SaskHealth)) after April 20, 2021, and before October 20, 2021, excluding tweets that were replies to other Twitter users.

Suppose in the study that the selected provincial governments only post tweets during operating hours in a day (e.g. 8:30 a.m. to 6:00 p.m.). In this case, one possible source of study error related to the variate *time.of.day* is the average number of tweets posted between 6:00 p.m. and 10:00 a.m. in the target population may differ from the study population.

The sampling protocol of this study involved using a browser-based tweet downloader that will download a sample of tweets from a 'primary' dataset. Data are taken from eight Canadian provincial health authority Twitter accounts: Alberta (@GoAHealth), British Columbia (@PHSAofBC), Newfoundland and Labrador (@HCS GovNL), Nova Scotia (@HealthNS), Ontario (@ONThealth), Prince Edward Island (@Health PEI), Quebec (@sante qc), and Saskatchewan (@SaskHealth). The dataset contains all tweets published by the eight accounts on or after April 20, 2021, and before October 20, 2021, excluding tweets that were replies to other Twitter users. The tweets were downloaded using the rtweet package in R on October 25, 2021. The sample size of my dataset is 923. 37 are sampled from the Prince Edward Island accounts and 47 from the Quebec accounts, 142 tweets are sampled from the British Columbia, 149 from the Nova Scotia, and 100 from the Saskatchewan accounts, and 170 tweets are sampled from the Alberta, 115 from the Newfoundland and Labrador, and 163 from the Ontario accounts.

One possible source of measurement error related to the variate *long.words* is the number of long words in a tweet is measured incorrectly.

Data

Some data that are collected may be less valuable than others; that is, some variates are not as important as other variate. The app that generates the data and the retweet package may record the data incorrectly.

Analysis 1

The variate *url.binary* is a binary variate, so binomial model is a suitable model for this variate since for any units, the variate *url.binary* for this unit is either 1 or 0. The parameter θ corresponds to the proportion of tweets in the target population that contains at least one url. Out of 923 samples, 372 tweets contain at least one url. The maximum likelihood estimate of θ is $\frac{372}{923} = 0.4030$.

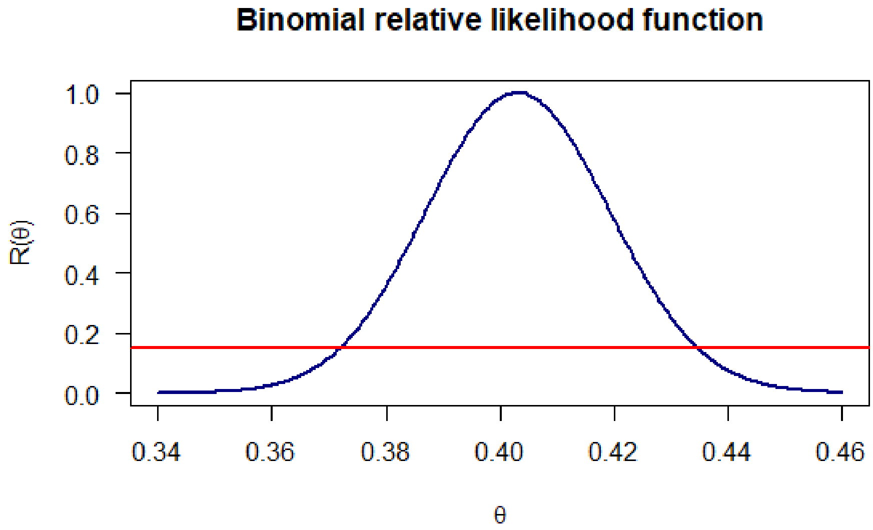


Figure 1 Relative likelihood function for the proportion of tweets in the study population that contain at least one url

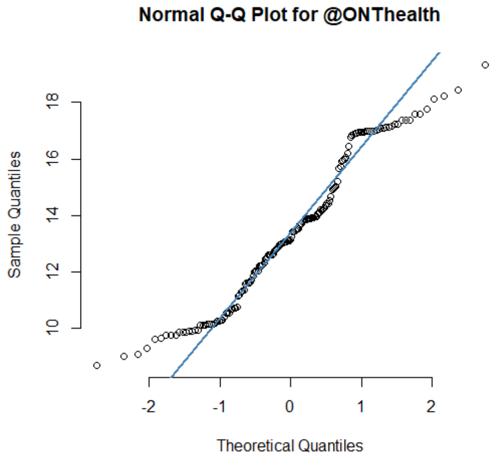
The 15% likelihood interval for the proportion of tweets in the study population that contain at least one url can be obtained approximately from Figure 1 by determining the two points of intersection of relative likelihood curve for the study(blue curve) and the red line R(θ) = 0.15. The 15% likelihood interval is [0.3718657, 0.4348000].

A recent study estimated that approximately 19% of tweets on Twitter contained urls. By the invariance property of maximum likelihood estimate, the observed value of the likelihood ratio test statistic $\lambda = \lambda(\theta_0)$ for testing the hypothesis $H_0: \theta = 0.19$ (which is a function of the maximum likelihood estimate of θ) is 223.183. The approximate distribution of the test statistic is binomial distribution. The p-value is extremely close to 0 which indicates that there is very strong evidence against the null hypothesis.

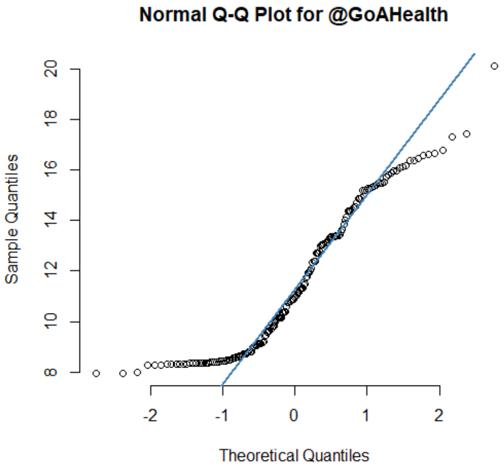
Analysis 2

If a $G(\mu_O, \sigma_O)$ model is assumed for the variate time.of.day.hour for tweets for @ONThealth, the parameter μ_O correspond to the mean tweet time in a day of tweets for @ONThealth in the study population. If a $G(\mu_A, \sigma_A)$ model is assumed for the variate time.of.day.hour for tweets for @GoAhealth, the parameter μ_A correspond to the mean tweet time in a day of tweets for @GoAHealt in the study population.

If Gaussian model fits the data well, the sample skewness should be close to 0, and the sample kurtosis should be close to 3. However, the sample skewness for the data for @ONThealth is 0.2029, and the sample kurtosis is 2.0176 which indicates that a Gaussian model is not a good fit for the data. The Normal qq-plot for the data for @ONThealth is given below. The plot shows that the distribution of the data seems bimodal.



The sample skewness for the data for @GoAHealth is 0.4737662, and the sample kurtosis is 2.054049 which also indicates the Gaussian model is not a good fit for the data. The Normal qq-plot for the data for @GoAHealth is given below. The S-shaped pattern of the plot shows that a uniform distribution may be a better model for this set of data.



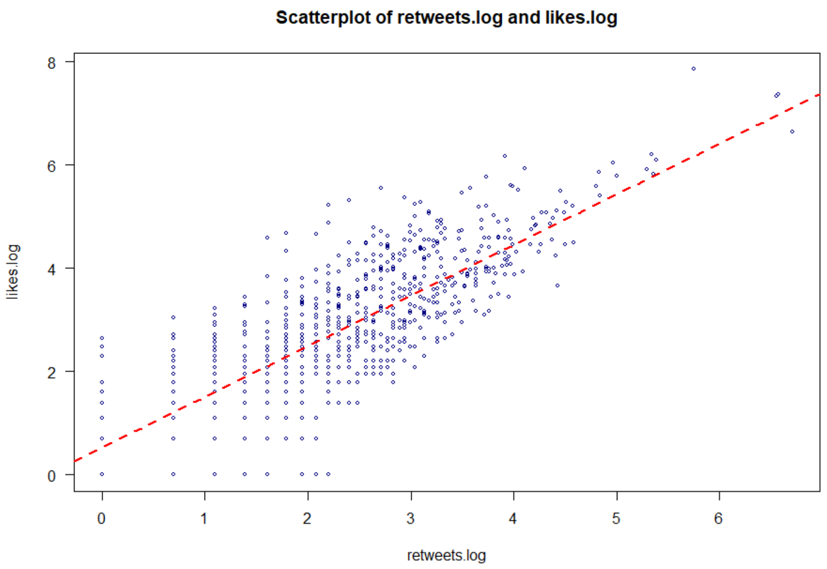
A 99% confidence interval for σ_O is [2.21, 2.93]. A 99% confidence for σ_A is [2.52, 3.33]. Base on these intervals it is not reasonable to assume $\sigma_A = \sigma_O$ because the overlapping region of these intervals is small.

The point estimate of the mean difference for the test of hypothesis $H_0: \mu_O = \mu_A$ is 1.845. A 95% confidence interval for $\mu_O - \mu_A$ is [1.26, 2.43]. The p-value 6.82×10^{-10}. The assumption of the independence of two variables is made.

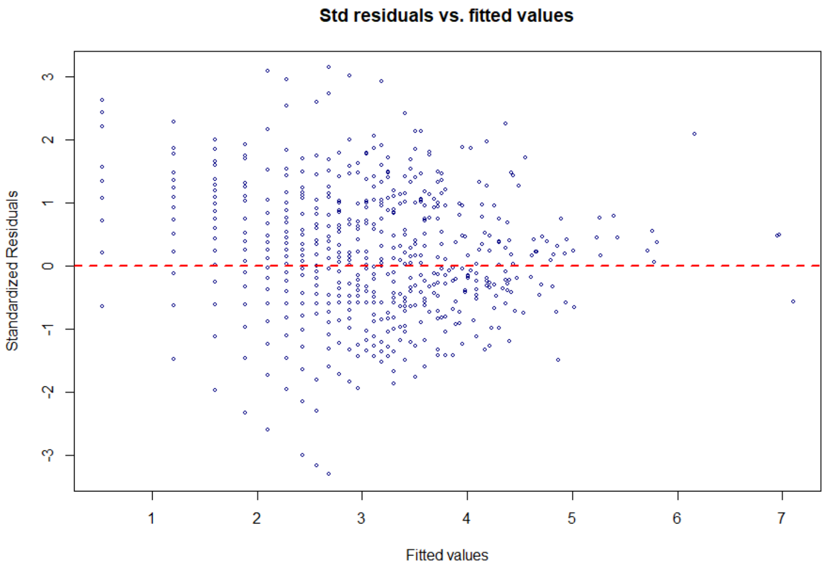
The meaning of the hypothesis $H_0: \mu_O = \mu_A$ is @ONThealth and @GoAHealth have the same average time of sending tweets in a day. Based on the p-value, we can conclude that there is strong evidence against the hypothesis. That is, there is strong evidence support that these data have different mean.

Analysis 3

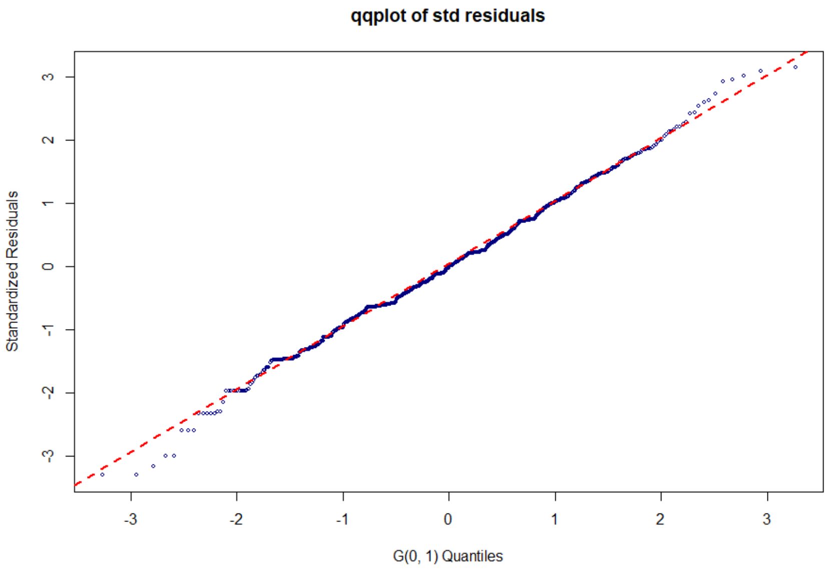
If a simple linear regression model $y = \alpha + \beta x$ is fitted to the explanatory variate x = retweets.log and the response variate $y =$ likes.log The least square estimates of α and β are 0.52 and 0.98 respectively.



The scatterplot checks whether the response variate likes.log and be modeled by a random variable whose mean is a linear function of the explanatory variate retweets.log and whose standard deviation is constant over the range of values of retweets.log. In this case, the observed points lie reasonably along the fitted line while the variability about the fitted line varies which indicates that the variance is not constant.



The standard residual checks the assumption about the form of the mean. If a simple linear regression model fits the data very well, points should lie roughly within a horizontal band of the fitted line. For this particular data, the plot has a heteroscedastic behavior. That is, the variance is not constant.



This plot checks the normality of assumption, the qq-plot gives approximately a straight line, so the normality assumption holds.

The parameter β corresponds to the change in mean $\mu = \alpha + \beta x$ for an unit increase in the explanatory variate retweets.log in the study population.

A 95% confidence interval for the parameter β is [0.9358, 1.0262]. $\beta = 1$ is in the 95% confidence interval. We can conclude that the p-value is greater or equal to 1-0.95 = 0.05.

The hypothesis $H_0: \beta = 1$ might be of interest because the maximum likelihood estimate of β is close to 1, and $\beta = 1$ indicates that an unit increase in the explanatory variate retweets.log in the study population will result in a unit increase in the response variate likes.log.

For a randomly chosen tweet in the study population with 30 retweets, a point estimate for the number of likes is 49.0319, and a 90% prediction interval for number of likes is $[e^{2.5554} - 1,\ e^{5.2295} - 1]$ = [11.8766, 185.7042].

Analysis 4

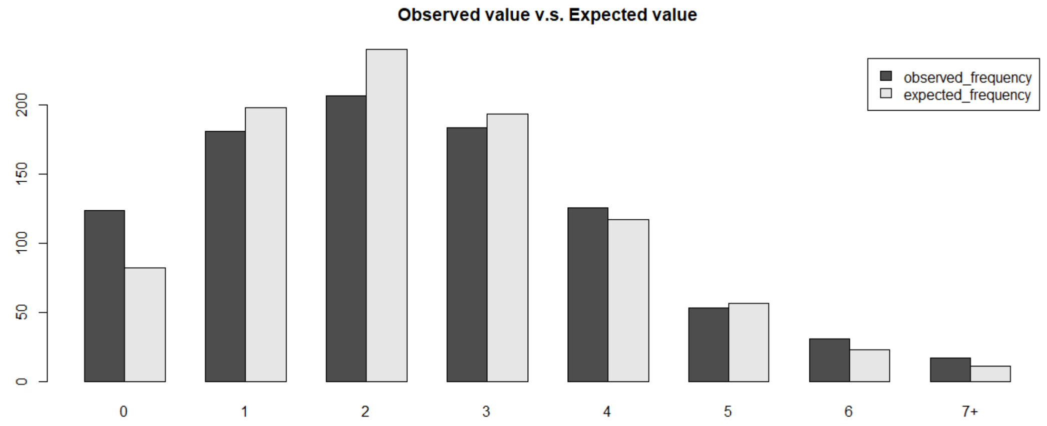
Assume a Poisson(θ) model for the variate long.words, a maximum likelihood estimate of θ is 2.420. An approximate 95% confidence interval for θ based on the asymptotic Gaussian pivotal quantity is [2.3200, 2.52073].

Number of long.words in a tweet	Observed Frequency	Expected Frequency
0	124	82.0444
1	181	198.5777
2	207	240.3156
3	184	193.8841
4	126	117.3177
5	53	56.7904
6	31	22.9089
7+	17	11.1710

Our model is Multinomial$(n; \theta_0, \theta_1, \dots, \theta_7)$, and $\Sigma_{j=0}^{7}\theta_j = 1$, which is a function of 7 parameters. H_0: Data fit a Poisson model or, more specifically $H_0\colon \theta_j = \frac{\theta^j e^{-\theta}}{j!}, j = 0,1,2, \dots ,6$

Under H_0 there is one unknown parameter θ which must be estimated. Therefore, the degrees of freedom for the Chi-squared approximation for the likelihood ratio test statistic are 7-1=6. The expected frequencies are all at least five, so we can use Chi-squared approximation to obtain the p-value.

The observed value of the likelihood ratio statistic is 31.5682, the p-value is 0.00001974.



For a randomly chosen tweet in the study population, a point estimate of the probability that this tweet contains at least 2 long words is 0.6696.

Analysis 5

To test the hypothesis of independence between media.binary and hashtag.binary using the likelihood ratio test statistic, the table of observed frequency is given below:

hashtag.binary/media.binary Observed	Contains media item	Doesn't contain media item	total
Contains hashtags	162	172	334
Doesn't contain hashtags	175	414	589
total	337	586	923

The expected frequency table is calculated:

hashtag.binary/media.binary Expected	Contains media item	Doesn't contain media item	total
Contains hashtags	$\frac{337\times334}{923} = 121.948$	$\frac{334\times586}{923}$ =212.052	334
Doesn't contain hashtags	$\frac{337\times589}{923}$ =215.052	$\frac{586\times589}{923}$ =373.948	589
total	337	586	923

The likelihood ratio test statistic

$$\lambda = 2\left[162\log\left(\frac{162}{121.948}\right) + 172\log\left(\frac{172}{212.052}\right) + 175\log\left(\frac{175}{215.02}\right) + 414\log\left(\frac{414}{373.948}\right)\right] = 32.1722$$

the approximate distribution for the test statistic Chi-square distribution with degree of freedom equal to 1.

The p-value is approximately equal to $P(W \geq 32.1722) = 1.41 \times 10^{-8}$, where $W \sim \chi^2(1)$.

The meaning of the hypothesis of independence in the context of this study is, for tweets in the study population, whether one tweet has hashtags is not affected by whether it has media items. That is, having hashtags and having media items are independent for tweets in the study population. There is very strong evidence against the hypothesis of independence. There is very strong evidence against the two variates have no relationship.

Analysis 6

Let Y_j be the number of tweets which are retweets from other accounts for provincial health agency i, $i = 1,2 \dots,8$. Assume that Y_i has a Binomial(n_j, θ_j) distribution, $j = 1,2,\dots,8$ where n_j is the number of tweets sampled from provincial health agency j and θ_j is the probability a randomly chosen tweet from the tweets in the study population for provincial health agency j is a retweet.

Assume the order @GoAHealth, @HCS_GovNL, @Health_PEI, @HealthNS, @ONThealth, @PHSAofBC, @sante_qc, @SaskHealth for the health agencies.

The meaning of the hypothesis $H_0 : \theta_1 = \theta_2 = \cdots = \theta_8$ is the proportion of tweets that are retweets in all tweets sent by the 8 accounts in the study population are equal.

The maximum likelihood estimate of $\theta_1, \theta_2, \dots, \theta_8$ are 0.888, 0.809, 0.0270, 0.322, 0.828, 0.5, 0.0426, 0.21. If we assume $: \theta_1 = \theta_2 = \cdots = \theta_8 = \theta$, then the maximum likelihood estimate for θ is 0.5655.

The table of observed and expected value are given below (the expected value are in parenthesis):

	Number of tweets that are retweets	Number of tweets that are not retweets	**Total**
@GoAHealth	151 (96.14301)	19 (73.85699)	170
@HCS_GovNL	93 (65.03792)	22 (49.96208)	115
@Health_PEI	1 (20.92524)	36 (16.07476)	37
@HealthNS	48 (84.26652)	101 (64.73348)	149
@ONThealth	135 (92.18418)	28 (70.81582)	163
@PHSAofBC	71 (80.30769)	71 (61.69231)	142
@sante_qc	2 (26.58072)	45 (20.41928)	47
@SaskHealth	21 (56.55471)	79 (43.44529)	100
Total	522	401	923

The observed value of the Pearson's Chi-squared Goodness of Fit test statistic is 662.6819, the approximate distribution for the test statistic Chi-square distribution with degree of freedom equal to 7. The p-value is extremely close to 0.

Conclusion

In the target population what proportion of tweets contain at least one url?

40.3% of tweets in the study population contain at least one url which is the maximum likelihood estimate of the sample proportion. The uncertainty in this estimate is reasonably small since the 15% likelihood interval is [0.3718657, 0.4348000] which is quite narrow.

Does the Gaussian model fit the variate time.of.day for @ONThealth in the target population? In the target population, is the mean time of sending tweets in a day for @ONThealth the same as @GoAhealth?

Gaussian model is not a good fit of the variate time.of.day for @ONThealth in the study population. The sample skewness for the data for @ONThealth is 0.2029, and the sample kurtosis is 2.0176 which is not close to 3. A test of hypothesis was conducted and the p-value for the hypothesis: @ONThealth and @GoAhealth have the same mean time of sending tweets in a day is 6.82×10^{-10} which indicates that there is strong evidence against the hypothesis. The 95% confidence interval for the mean difference is [1.26, 2.43] and 0 is not in this interval.

Suppose a tweet with 30 retweets is drawn from the target population. What is the approximate number of likes it will get? In the target population, if the logged number of a retweet is increased by one, how will the number of like changes?

In the study population, a point estimate of the number of likes that a randomly chosen tweet with 30 retweets is 49.0319. A 90% prediction interval is [11.8766, 185.7042]. The uncertainty in this estimate is large since the 90% prediction interval is quite wide. In the study population, a unit increase in logged number of retweets will result in 0.98 unit of logged number of likes. A 95% confidence interval for this parameter is [0.9358, 1.0262]. The uncertainty in this estimate is small since the 95% prediction interval is quite narrow.

Suppose a tweet is randomly chosen from the target population, what is the probability that this tweet contains at least 2 long words?

An estimate of the probability that a tweet, chosen at random from the study population, contains at least 2 long words is 0.4453. The uncertainty in the estimate is reasonably small since the sample size for this study is n = 923 is large.

Are the tweets in the target population that contain hashtags more likely to contain or do not contain media item?

A likelihood ratio test of hypothesis for the independence between the two variates in the study population was conducted, the likelihood ratio test statistic is 32.1722 and the p-value is 1.41×10^{-8}. This indicates that there is strong evidence against the hypothesis that the two variates have no relationship.

In the target population, do the 8 different accounts have the same proportion of tweets that they sent are retweets from other accounts?

In the study population, suppose all accounts' is.retweet variate satisfies binomial distribution. A two way table and likelihood ratio test of hypothesis was constructed to test H_0: the proportion of tweets that are retweets in all tweets sent by the 8 accounts in the study population are equal. The observed value of the Pearson's Chi-squared Goodness of Fit test statistic is 662.6819, and the p-value is extremely close to 0. So there is strong evidence against H_0.

One limitation of the study is that the number of samples for different provincial health agencies is different. For example, only 37 tweets were sampled from @Health PEI. Results may be inaccurate based on insufficient sample size.

Time constraint is also a limitation of the study. As more and more people get vaccinated, the pandemic will get controlled gradually. The data collected between April 20, 2021, and October 20, 2021, generally reveals the media strategy on Twitter in the past. It may not be very reflective of media strategies from 1 January 2022 to 31 March 2022.