Midterm Report

Problem part:

This report aims at helping the Canadian government to anticipate the Twitter activities of provincial health agencies during the period from 1st January 2022 to 31st March 2022. So we define the set of all tweets published by all provincial health agencies during the period from 1st January 2022to 31st March 2022 in Canada as the target population, and the unit is a tweet published by any provincial health agency during the period from 1st January 2022 to 31st March 2022 in Canada.

In this study, the data of tweets posted by those official accounts of provincial health agencies includes variates: 'health', 'covid', 'vaccine'(indicates whether the tweet contains the corresponding keywords), 'media'(the number of media items likes images or videos in tweet), 'first.tweet'(whether the tweet is the first one of the day), 'retweets'(number of users who have retweeted a tweet), 'time.of.day'(the time of day the tweet was published), 'likes'(number of users who have liked a tweet), 'username'(the username of the account that sent the tweet) and 'is.retweet'(whether the tweet is a retweet of another account). Among these variates, 'health', 'covid', 'vaccine', 'first.tweet' and 'is.retweet' are all binary variates, which are taking value of 1 for yes, and 0 for otherwise. We are interested in the proportion of tweets contain a particular keywords('health', 'vaccine', or 'covid') among all tweets published by provincial health agencies during that specific period, the mean number of media items in the first tweet of a day among all tweets published by provincial health agencies during that specific period, and the proportion of retweets from other accounts among all tweets posted by provincial health agencies during that specific period. The variate 'media', 'retweets' and 'likes' are discrete variates. In the following study, we will focus on the mean number of media items used, retweets and likes received among all tweets published by provincial health agencies during the specific period separately. 'time.of.day' is a continuous variate and 'username' is a categorical variate. Meanwhile, the mean time of day that tweet was published among all tweets published by provincial health agencies during the specific period and the proportion of tweets sent by each username's account among all tweets posted by provincial health agencies during the specific period will help us draw the conclusion of interest.

For a comprehensive analysis, we include a series of motivating questions with different types: question (a), (c) to (h) are descriptive problems, and question (b) is a predictive problem.

Plan part:

Since we only observed and recorded all distinct variates of tweets posted by provincial health agencies during that specific period without controlling any of them, the study is observational. Our study population is a total of 21,883 tweets stored in a 'primary' dataset and posted by eight Canadian provincial health official accounts, which are Alberta (@GoAHealth), British Columbia (@PHSA of BC), Newfoundland and Labrador (@HCS-FovNL), Nova Scotia (@HealthNS), and Saskatchewan (SaskHealth), on or after 20th April 2021, and before 20th October 2021. Notice that all tweets that were replies to other accounts were excluded.

For consideration of time and costs, the 'primary' dataset is the most appropriate study population that we can use. However, there still exists some restriction which may lead to study error. We want to predict the tweet activities during the period from 1st January 2022 to 31st March 2022, while all tweets in the study population are from 20th April 2021 to 20th October 2021. It is likely that the pandemic will close to the end in the first half of next year, and there will be less information about COVID-19 and vaccine need be released to the public. So the proportion of tweets contains keywords of 'covid' or 'vaccine' might be much higher in the study population than in the target population. Besides, our sample population does not include the accounts of Nunavut, Yukon and North west Territories, which might causes the overestimation of the mean number of likes received, since there are less population in the three regions and thus less people follow the tweets posted by the three regions' health agencies.

To define the sampling protocol, we access a browser-based tweet downloader and entered the passwords and ID number, then the downloader generated a sample of tweets from each of the eight provincial health accounts from the 'primary' dataset randomly. There is a total of 982 tweets in our sample. Some possible source of measurement error also exists in our study. For the keyword variate 'covid', since the variate was measured by evaluating whether the particular string of characters appears or not, it is possible that some words might contain this particular string of characters but not refer to the exact 'covid' that we want. For example, 'covidien' is the name of a global manufacturer of medical devices and supplies, and the word is not refer to 'covid' that we are of interest. However, the word will be identified as a keyword because it contains the string of 'covid', which will lead to measurement error.

Data part:

Our sample was generated by a browser-based tweet downloader randomly. For the account of Newfoundland and Labrador's health agency (@HCS-GovNL), the number of retweets are between 0 to 200 in most cases, but there is a tweet with 1415 retweets on 26th April 2021, which is extremely high and need further analysis on it. We should notice that some important announcement is very likely to attract public's attention and leads to extreme value in some varieties, such as 'likes' and 'retweets'.

Analysis Part:

We are looking forward to help the government formulate social media strategy on what material to tweet. We also believe that tweets about COVID-19 are more likely to catch public's attention during the pandemic, so we choose 'covid' as the most important keyword for the purpose of this study. It is obvious that the variate 'covid' has only 2 types of distinct outcomes: contains 'covid' or not; each time tweet contains 'covid' will not affect the probability that other tweet contains this keyword, so the experiment is repeated independently and the probability that tweet contains 'covid' is fixed every time. Therefore, the binomial distribution is an appropriate model to estimate the proportion of tweets with keywords 'covid' among the total of 21,883 tweets stored in the 'primary' dataset. And this estimate can be obtained through the observed number of tweets containing 'covid' dividing by the sample size, which is actually $\frac{494}{982} \approx 0.503$. The 15% likelihood interval estimate for the proportion of tweets in the study population that contain 'covid' is [0.4720045, 0.5340888]. Now we suppose 50 tweets are chosen at random from the total of 21,883 tweets, let Y be the number of tweets containing 'covid', so Y ~ approximately G(μ, σ) with $\mu = 50 \times \frac{494}{982}$,$\sigma = \sqrt{50 \times \frac{494}{982} \times (1 - \frac{494}{982})}$ according to central limit theorem, and we use P($Y \geq 25$) = 1-P($Y \leq 24$), given $\frac{Y-\mu}{\sigma} \sim G(0,1)$ to calculate the probability that at least 25 tweets contain 'covid', the result is 0.5732399.

For variate 'media' and 'first.tweet', the following table demonstrates the number of media items used for first tweets of the day and tweets which are not first tweets of the day.

	0	1	2	3	4	Total
First tweet	68	77	0	0	1	146
Not first tweet	558	261	11	3	3	836
Total	626	338	11	3	4	982

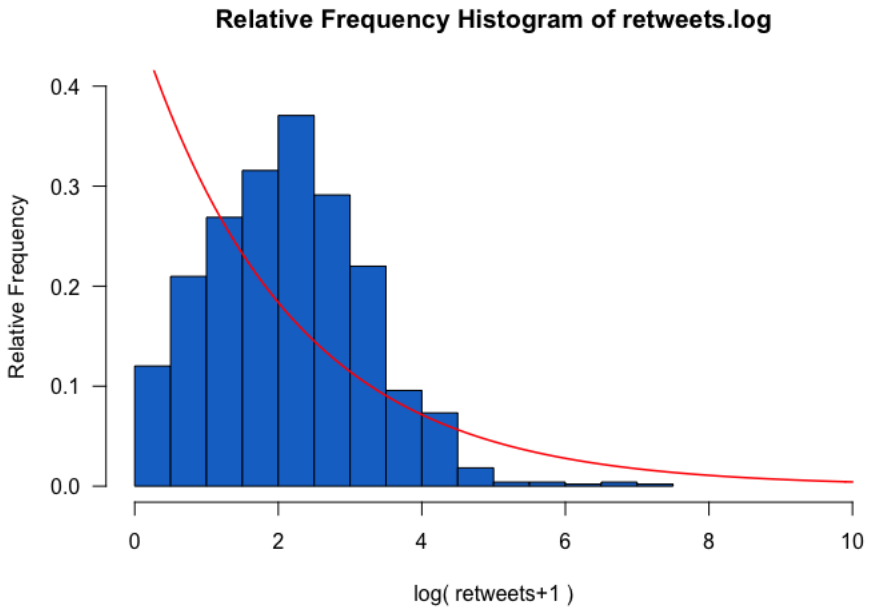
The sample mode for the variate 'media' for first tweets of the day and the variate 'media' for tweets which are not first tweets of the day are both 0. The sample mean for them are 0.5547945 and 0.3636364 separately.

We found that the number of media items used for first tweets of the day is non-overlapping intervals are independent. For sufficiently shout period, the probability of 2 or more first tweet of the day are published in the interval is very close to zero. And the number of media items used in first tweets of the day occurs at a homogeneous rate. According to above, we believe that Poisson distribution could be used for modeling 'media'. And the only one unknown parameter in this model is the average number of media items used for a first tweet of the day. Furthermore, the maximum likelihood estimate for the unknown parameters can be obtained through the total number of media items used in all first tweets of the day dividing by the total number of first tweets of the day, which is $\frac{81}{146} \approx 0.55479$.

In order to analyze the variate 'retweets', we create the transformed variate 'retweets.log'= log(retweets+1). Following is the table of five number summary and skewness for the variate 'retweets.log'.

	x(0)	q(0.25)	q(0.5)	q(0.75)	x(n)	Sample skewness
retweets.log	0.00	1.39	2.20	2.89	7.26	0.29

Now, we assume that retweets.log follows and Exponential (λ) distribution, then the maximum likelihood estimate of parameter λ is 2.120346. We note that the parameter λ is the average of log (number of tetweets+1) for one tweet among the total of 21,883 tweets published by provincial health agencies. We also provide the plot of the relative frequency histogram of retweets.log combined with Exponential ($\hat{\lambda}$) probability distribution function.

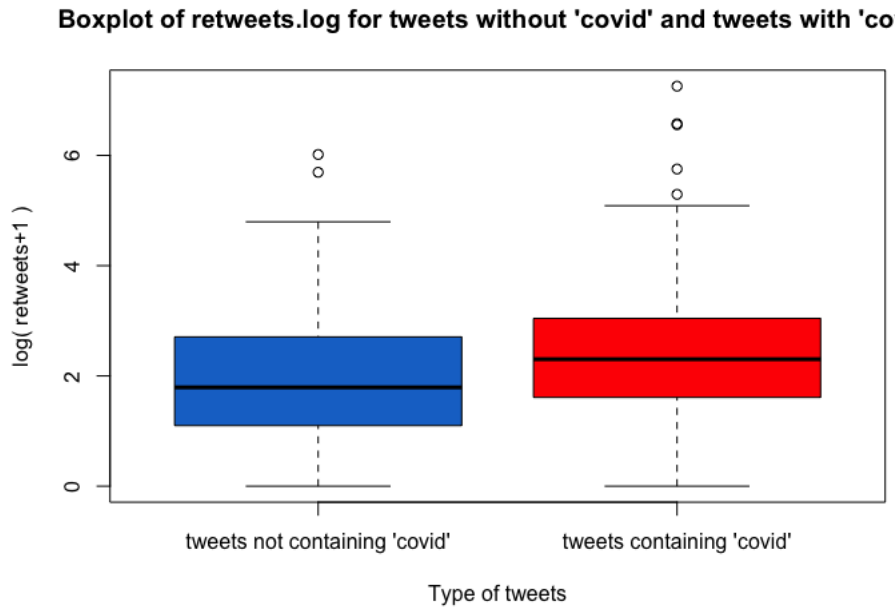


In further, the variate 'retweets.log' related to the keywords 'covid' will be examined in the following. We provide the five-number summaries for both the number of retweets for tweets containing 'covid' and the number of retweets for tweets not containing 'covid'.

	x(0)	q(0.25)	q(0.5)	q(0.75)	x(n)
Number of retweets with 'covid'	0.00	4.00	9.00	20.00	1415.00

	x(0)	q(0.25)	q(0.5)	q(0.75)	x(n)
Number of retweets without 'covid	0.00	2.00	5.00	14.00	409.00

Now, we assume retweets.log follows an Exponential(λ_0) distribution for tweets not containing 'covid' and an Exponential(λ_1) distribution for tweets containing 'covid'. Hence, the maximum likelihood estimate of λ_0 is 1.931229, the maximum likelihood estimate of λ_1 is 2.307166. Besides, the 95% confidence interval for λ_0 based on asymptoic Gaussian pivotal quantity is [1.759884, 2.102574]. The 95% confidence interval for λ_1 based on asymptoic Gaussian pivotal quantity is [2.103713, 2.510619]. A side-by-side boxplot of retweet.log for tweets not containing 'covid' and containing 'covid' is shown as below:



According to this plot, the range of 'retweets.log' for tweets containing 'covid' is a bit wider than the range for tweets not containing 'covid'. While, the IQR of 'retweets.log' for tweets containing 'covid' is smaller than the IQR for tweets without keyword 'covid'. Besides, the location of the median line of 'retweets.log' for tweets containing 'covid' is higher than the one for tweets not containing 'covid'.

Now, we start to analyze the variates 'likes', and 'time.of.day'. For convenience, we create a new variate which converts time.of.day into hours. The five-number summaries for the number of likes received by tweets which were tweeted during the period 9:00-12:00 is shown as below:

	x(0)	q(0.25)	q(0.5)	q(0.75)	x(n)
Number of likes received by tweets during 9:00-12:00	0.00	5.00	14.00	48.25	3706.00

The five-number summaries for the number of likes received by tweets which were tweeted during the period 12:00-15:00 is shown as below:

	x(0)	q(0.25)	q(0.5)	q(0.75)	x(n)
Number of likes received by tweets during 12:00-15:00	0.00	3.00	9.00	22.00	1401.00

Following the request, we apply a Gaussian model for the number of likes received by tweets which were tweeted during 9:00-12:00, which is $G(\mu_{am}, \sigma_{am})$, and $G(\mu_{pm}, \sigma_{pm})$ for the number

of likes received by tweets which were tweeted during 12:00-15:00. So the maximum likelihood estimates of μ_{am}, $\mu_{pm}$is 57.67808 and 28.50462 separately. The maximum likelihood estimate of σ_{am}, σ_{pm} is 268.7647 and 108.7054 separately. The 95% confidence interval for μ_{am} is [26.72251, 88.63365], the 95% confidence interval for μ_{pm} is [16.64194, 40.3673]. The 90% confidence interval for σ_{am} is [251.7038 , 288.5351], the 90% confidence interval for σ_{pm} is [102.142, 116.2527].

We proposed a binomial distribution to examine how often different provincial health agencies use information retweeted from other accounts. Since there are two distinct outcomes for is.retweets: using information retweeted form other accounts or not. Besides, each time the provincial health agencies using retweets or not will not affect the probability that other tweets use retweeted information, so the trials are repeated independently. The parameter of this model is the probability that a provincial health agencies uses information retweeted for other accounts.

For comparison, we list the maximum likelihood estimate and 95% confidence interval of the parameter for distinct provincial health agencies.

For Alberta, the maximum likelihood estimate of the probability of using retweets is $\frac{164}{179} \approx 0.9162$, the 95% confidence interval is [0.8756095, 0.95679277]

For British Columbia, the maximum likelihood estimate of the probability of using retweets is $\frac{57}{124} \approx 0.4597$, the 95% confidence interval is [0.3719591, 0.5473958]

For Newfoundland and Labrador, the maximum likelihood estimate of the probability of using retweets is $\frac{155}{200} \approx 0.775$, the 95% confidence interval is [0.7171271, 0.8328729]

For Nova Scotia, the maximum likelihood estimate of the probability of using retweets is $\frac{31}{117} \approx 0.26496$, the 95% confidence interval is [0.1849924, 0.3449222]

For Prince Edward Island, the maximum likelihood estimate of the probability of using retweets is $\frac{1}{45} \approx 0.02222$, the 95% confidence interval is [-0.02084587, 0.06529032]

For Ontario, the maximum likelihood estimate of the probability of using retweets is $\frac{122}{149} \approx 0.818792$, the 95% confidence interval is [0.7569433, 0.8806406]

For Quebec, the maximum likelihood estimate of the probability of using retweets is $\frac{1}{40} = 0.025$,

the 95% confidence interval is [-0.02338273, 0.07338273]

For Saskatchewan, the maximum likelihood estimate of the probability of using retweets is $\frac{31}{128} \approx 0.2422$, the 95% confidence interval is [0.1679711, 0.3164039]

Conclusion part

Some conclusions targeted our motivating question at the beginning of the report can be drawn according to our data analysis.

(a) The proportion of tweets contain a keyword 'covid' is estimated to be 0.503 in the total of 21,883 tweets posted by eight Canadian provincial health agencies during the period from April 20,2021 to October 20,2021. And stored in 'primary' dataset.

(b) Suppose 50 tweets are drawn at random from the total of 21,883 tweets stored in 'primary' dataset, the probability that at least 25 tweets will contain 'covid' is estimated to be 0.573.

(c) We expect that most of first tweets of the day and not first tweets of the day contain 0 median items, while the mean number of media items used for first tweets among the total of 21,883 tweets during the specific period is expected to be 0.55479, compared with 0.36364 for not first tweets.

(d) We also expect to see the mean number of retweets in the 21,883 tweets published by provincial health agencies during the specific period is 17.3544.

(e) According to the relative frequency histogram of retweets.log with a superimposed exponential probability distribution function, we found there exists big difference between the relative frequency histogram and exponential probability distribution function, the relative frequency histogram of retweets.log seems more symmetric than we expected, so the exponential distribution is not a good model for the transformed variate, retweets.log for the total of 21,883 tweets.

(f) After analyzing the number of retweets in our sample with and without keywords 'covid' by side-by-side no plot and five-number summary table, we estimated that the distribution of retweets for tweets with keywords 'covid' has wider range and higher median than the distribution of retweets for tweets without 'covid' among the total of 21,883 tweets, which means that the distribution of retweets for tweets with 'covid' intends to be higher than the distribution of the one without 'covid' generally.

(g) In the total of 21,883 tweets published by provincial health agencies during the period from

April 20,2021 to October 20;2021, which is our study population, we expect that the mean number of likes received during 9:00-12:00 is 57.678, which is higher than 28.5046 for likes received during 12:00-15:00. It seems that sending a tweet during 9:00-12:00 is likely to receive more likes than sending during 12:00-15:00 in afternoon.

(h) After comparing the probability of using information retweeted from other accounts for distinct provincial health agencies, we found that for Alberta, Newfoundland and Labrador, Ontario, these provinces' health agencies use retweets much more frequently, which means the probability of using retweets is up to more than 75%. For British Columbia, the probability of using retweets is close to the probability of using original tweets. For Nova Scotia and Saskatchewan, the two provincial health agencies intends to create their original tweets with relative high frequency. Finally, for Quebec and Prince Edward Island, the probability of using retweets is very close to zero, which means retweet is almost never used.

Based on this study, we expect that there will be a considerable proportion of tweets posted by provincial health agencies is related to COVID-19 or vaccine. We also recommend that the health agencies can release more information about COVID-19, since the relative large number of likes and retweets demonstrates that the public are very concerned about the news on COVID-19. Besides, the period from 9:00 to 12:00 might be a good time for sending tweets.

As mentioned above, we should notice that there exists statistical uncertainty in answers related with variate 'covid', since some words includes the string of 'covid' ,such as 'Covidien'', might be counted as the existence of keyword. Besides, extreme values also bring us uncertainty about our result. There exists one to two tweets with extremely large number of likes and retweets, it is probably that some break news are released, which need further study.

There are also several limitations to our conclusions. Because our samples are all selected within the period from April 20,2021 to October 20,2021, the conclusion may not apply to the tweets which will be posted from January 1,2022 to March 31,2022, it is possible that the pandemic is coming to the end in the near future, and less tweets about COVID-19 or vaccine will be published by health agencies in 2022, so some distribution of variate such as the proportion of tweets containing 'covid' will be actually lower than our estimation. In this study, our analysis is only limited to tweets which are not replies to other Twitter user, so the conclusion might be not suitable for tweets which are replies to other user among all tweets published by provincial health agencies.

In most cases, the tweets which are replies to other user gain less likes and retweets than usual tweets. Finally, there are eight provincial health agencies are included in our study, so we also worry that the tweets posted by other provincial health agencies, such as Nunavut, Yukon and Northwest Territories are largely different with the eight provinces in our study due to the much smaller population and less information about pandemic need to be posted in these regions. Thus it might be not appropriate to apply our conclusions to all provincial health agencies.