Understanding Canadian Smokers

Assignment 1 - STA304 - Fall 2021

Dhanraj Patel - 1003965168

Contents

Part 1 - Inroducing the Survey

Goal

For many years now after extensive scientific research and awareness campaigns Canadians have become increasingly aware about the different types of cancers and diseases associated with the use of cigarettes. Due to the increase in education, the amount of smokers in Canada has been steadily decreasing since 1955 with the amount of smokers observed in recent years showing the lowest smoking rates ever observed [1]. However, despite being historically low in levels, a recent Statistics Canada survey records that there are still over 4 million Canadians who continue to smoke cigarettes. [2]. The goal of this survey is to gain a better understanding of the smoking habits and reasons as to why those who smoke continue to do so despite knowing the risks associated. By analyzing the data in the survey, we are able to utilize the responses in order to create more effective campaigns and programs that help Canadians quit smoking.

Procedure

The purpose of the survey is to gain a better understanding of the habits and reasons as to why Canadian smokers continue to smoke, in order to do this the sample population needs to be diverse and representative of the entire Canadian smoking population. In order to implement this survey, one proposed method is to conduct the survey in online Canadian smokers help forums such the smokers help line forum [3]. These forums are specifically made to act as a support groups and discussion forums where Canadians smokers can congregate and discuss their smoking habits. By conducting our survey in these forums we are able to reach a large and diverse set of Canadian smokers from all provinces and territories, which helps us to reduce sampling bias. In addition, it is important to make our survey as anonymous as possible while still trying to analyze the habits and reasons people continue to smoke in order to limit the non-response bias in the survey. Within the survey the target populations is the population of Canadians who smoke cigarettes. However, since the proposed survey procedure will conduct the survey in online Canadian smokers forums, the frame population is all the Canadian smokers who visit the forums where the survey is being conducted. Lastly, the sample population will be those in the Canadian smokers forums who decide to participate in the survey.

Some of the advantages of conducting the survey in these online forums is that the survey is able to reach a large and diverse amount of Canadian smokers. Since these forums are visited by tens of thousands of Canadians, this method will allow us to mitigate the sampling bias in our analysis and help us understand the habits of a wide array of Canadian smokers. However, there are drawbacks in using this method which include the fact that this survey will exclude those who do not visit these forums. Our sample population would not include those who do not regularly use the internet and this may limit the ability to generalize of our survey. Furthermore, since the survey answers are voluntary and online where the participants are not vetted at all, this may lead to responses that may need to be vetted out.

Showcasing the survey

The full survey used in the study is available in the following link: https://forms.gle/mMKhpUn599uvfFxd9

Question 1

What is your most significant reason for smoking? (survey participant chooses one answer from the options)

- Addiction
- Relaxation
- Weight maintenance
- Helps stay focused
- Withdrawal symptoms
- Social activity with friends

Question 2

Do you intend to quit smoking in the next six months? (survey participant chooses one answer from the options)

- Yes
- No

Question 3

Approximately how many cigarettes do you smoke per day on average?

(please note that the survey participants will be entering a numerical short answer for this question, the answers for the question have been restricted to only accept numerical answers greater or equal to 0)

Significance of questions

The reason I chose these questions to display is that they altogether exemplify the central goal of survey of trying to better understand the habits of Canadian Smokers and the core reasons as to why they continue to smoke. Each of the questions has there individual strengths and draw backs.

As for the first question, the question aims to understand the most significant reasons as to why Canadian continue to smoke cigarettes. The six options to select where taken from a cancer.com article which listed the top six reasons people say as to why they smoke cigarettes [4]. The strengths of this question is that it allows the participants of the survey to select from a wide array of different options of reasons as to why they continue to smoke. The drawbacks of this question lies in the fact that the question asks them for their most significant reason and thereby limits them to selecting one option only. Due to this, participants are unable to select multiple factors that cause them to continue to smoke and instead focus on their most significant reason.

As for the second question, the question aims to to understand if the participants in the survey are actively attempted to quit their smoking habit. The question was based off an Ontario Tobacco Research Unit (OTRU) study which found that 60 percent of Canadian smokers intend to quit in the next six months [5]. The strengths of this question is that it
allows us to better understand if the survey participant is actively attempting to mitigate their smoking habit or not. The drawbacks of the question lies with the fixed time frame of six months used in the question. It does not different if the participant is attempting to quit in any time before the six months and leaves out participants who are intending to quit any time past the six months.

Lastly the third questions attempts to understand the smoking habits of the survey participants by asking how much cigarettes they approximately smoke in a day. The strengths of this question is that it asks the participants of the survey to manually enter in the average number of cigarettes smoked with a short answer question instead of choosing one answer from a multiple preset of answers. This will help get detailed and granular set of answers that can be analyzed. The drawbacks of this question is that it limits the participant into answering the average amount of cigarettes they smoke only in one day and does not differentiate their smoking patterns between the days of the week. It does not factor in if they only smoke on the weekdays, weekends or only at special events.

please note bibliography for part 1 and 2 can be found at bottom of report

Part 2 - Analyzing the Survey

Data

Explaining the simulation process

The main goal of the study is to analyze the habits of Canadian smokers to better understand their habits and reasons for smoking. For the Data section of this analysis, the data will be simulated. There will 500 survey partipants simulated in the dataset.

In order to simulate the average number of cigarettes smoked in a day for 500 people, I used a normal distribution with a mean of 13 and standard deviation of 1.75 cigarettes smoked per day on average. The mean and standard deviation were based of a 2017 university of Waterloo study that studied the habits of Canadian smokers which stated that the average number of cigarettes smoked in a day by Canadian smokers was 13 cigarettes [1]. In addition, the standard deviation utilized the `range over 4` rule [8], where the range of values was also based off the university of Waterloo study which used the range of values of the average number of cigarettes smoked in a day from 1999 to 2017. The normal distribution was used to account account for both the heavy smokers and those who seldom smoke only in social situations.

Furthermore, in order to simulate data for all 500 survey participants on if they intend to quit smoking in the next six months I used the discrete Bernoulli distribution ($bern(\theta)$), where the value of theta was 0.60. A value of 1 for the Bernoulli distribution would indicate that "yes" the survey participant did intent to quit smoking in the next six month and a value of 0 would indicate that "no" the survey participant does not intend to quit smoking in the next six months. The reason why I chose the Bernoulli distribution is that I wanted to measure a discrete binary statistic and that each trial is independent of each other. Lastly, I chose 0.60 for the value of theta is due an Ontario Tobacco Research Unit (OTRU) study which found that 60 percent of Canadian smokers intend to quit in the next six months [5].

In addition, In order to simulate if the 500 survey participants regularly use electronic cigarettes I used the discrete Bernoulli distribution ($bern(\theta)$) where theta is equal to 0.541. A value of 1 in each Bernoulli trial represents that the survey participant does regularly use electronic cigarettes and a value of 0 represents that the survey participant does not use regularly use electronic cigarettes. The reason I chose the discrete Bernoulli distribution is that wanted to simulate a binary statistic where each trial is independent of each other. Lastly, the value of theta which I chose to be 0.541 is based of an university of Waterloo study on Canadian cigarettes smoking habits and patterns where they found in their study that 54.1 percent of smokers also regularly used electronic cigarettes [1].

Moreover, in order to simulate for all 500 participants most significant reason for smoking reasons, I used a discrete uniform distribution with six possible options. The six options are addiction, relaxation, weight maintenance, helps stay focused, withdrawal symptoms and due to being a social activity with friends. These six reasons were based of a Canadian Cancer Society article which conducted a study which resulted in finding that the six aforementioned reasons as being the top six reasons Canadians smokers claimed to be the reason they smoke [4]. I choose to use a discrete uniform distribution due to the fact that in the article all six reasons were stated to be the top six reasons the participants continued to smoke in equal prominence [4] and the survey participant were asked to discretely choose from one of six options and each trial is independent from one another.

In order to simulate the gender of the 500 participants in the survey I used a discrete multinomial distribution in which each trial has a 60.965 percent chance of being male, 38.965 percent chance of being female and 0.07 percent chance of being non-binary. I chose to use a discrete multinomial distribution due to the fact that each trial is independent of one another and that in each trial, the simulation will discretely choose one of three options with different probabilities of being chosen. For the simulation, the probability of selecting the gender options were based of a 2020 Statistics Canada community health survey which found in their survey that 60.5 percent of smokers surveyed were male and 39.5 percent of smokers surveyed were female [4]. Since the health survey did not incorporate non-binary individuals, I used information gathered from the 2021 Canadian census that stated that 0.07 percent of Canadians identify as non-binary [6]. For my simulation I

used this metric in my simulation as the probability that the simulated trial would select non-binary as the survey participants gender and subtracted 0.035 from the probability of selecting both males and females to incorporate the 0.07 percent.

Also, in order to simulate the age interval of the simulated 500 participants I used discrete multinomial distribution. In each trial of the simulation, the simulation has a one percent chance of selecting the survey participant to be between the ages 12-17, 30 percent chance to be between 18-34, 26 percent chance to be between 35-49, 29 percent chance to be between 50-64 and a 14 percent chance to be above 64 years old. I chose to use a discrete multinomial distribution due to the fact that each trial is independent of one another and that in each trial, the simulation will discretely choose one of five options with different probabilities of being chosen. The probability of selecting the age intervals in the simulation is based of a 2020 Statistics Canada community health survey which investigated the characteristics and habits of Canadian smokers [4]. The simulation uses the age intervals and age proportions found in the Canadian community health survey.

Lastly, In order to simulate the number of friends the survey participant has that also smokes for all 500 simulated participants I used discrete multinomial distribution. In each trial of the simulation, the simulation has a 13.2 percent chance of selecting that the survey participant has zero friends who also smokes, 13.4 percent chance of having one friend who also smokes, 19.3 percent chance of having two friends who also smokes, 18.6 percent chance of having three friends who also smokes and 35.5 percent chance of having four or more friends who also smokes. I chose to use a discrete multinomial distribution due to the fact that each trial is independent of one another and that in each trial, the simulation will discretely choose one of five options with different probabilities of being chosen. The probability of choose the number of friends who also smoke was based of an U.S. National Institutes of Health's National Library of Medicine PMC article that aimed to study the relationship between the the number of friends who smoke and the habits of smokers. The study was conducted in the United States, Australia and Canada [7]. Within the simulation, the probability of selecting the number of friends each each simulated participant has that also smokes mirrors proportional finding of the Canadian portion of the study [7].

Data cleaning process

Due to the fact that the values for the dataset were simulated there were no instances of missing data to clean. Furthermore, since the data is designed to be anonymous, duplicate rows of data are equally valid and were not cleaned. Lastly, the data columns did not need to be modified in anyway to be used in the numerical summaries, graphical summaries or the analysis portion of the study.

Important variables

The following are the important variables of the dataset that were used for the numerical and graphical summaries in addition to the analysis portion of the study.

- Age - The age interval of the participant
- Gender - The gender of the survey participant
- Quitting_intentions - if the survey participant intends to quit smoking in the next six months.
- reasons_smoking - The most significant reason selected by the survey participant as why they continue to smoke.
- average_daily - The average number of cigarettes smoked in a day by the survey participants
- friends_smoke_count - The number of friends of the survey participants who also smoke cigarettes.
- ecig_use - If the survey participants regularly uses electronic cigarettes.

Numerical Summaries

After cleaning the data we can now move on and analyze important numerical summaries in our data set. Numerical summaries allow us to better understand the values collected in our data set. It is important to

analyze both the location and spread of our variables.

comparing gender to average number of cigarettes smoked in one day

Table 1: The trimmed mean is trimmed by 10 percent

gender	Count	Min	Q1	Median	Q3	Max	IQR	Mean	Trimmean	Var	SD	Range
Female	213	8.69	11.530	12.78	13.80	16.63	2.270	12.69	12.67	3	1.67	7.94
Male	287	7.60	11.685	12.99	14.24	17.14	2.555	12.84	12.90	3	1.72	9.54

In table_1 we see the numerical summaries of the average number of cigarettes smoked in one day separated for both males and female survey participants. From the count we see that the total number of males in the survey is greater than the number of females. In all quartiles, including the minimum value, q1, median, q3 and maximum value we see that the average number of cigarettes smoked by males is higher compared to the female values. In addition, the mean, trimmed mean and range of average number of cigarettes smoked in one day is also higher for males compared to females. From these results we can conclude that from the simulated results that, on average the males smoke grater amounts of cigarettes compared to females. Lastly, we see that from the 500 overall simulated survey participants the variance of both males and females is equal.

comparing age to average number of cigarettes smoked in one day

In addition to gender, we can also analyze the numerical summaries of the average number of cigarettes smoked in one day to the age intervals of the simulated survey participants.

Table 2: The trimmed mean is trimmed by 10 percent

age	Count	Min	Q1	Median	Q3	Max	IQR	Mean	Trimmean	Var	SD	Range
12-17	7	10.10	10.960	12.440	13.0000	13.46	2.0400	11.99	11.99	2	1.36	3.36
18-34	150	7.60	11.635	13.050	14.2625	16.63	2.6275	12.87	12.95	3	1.81	9.03
35-49	133	8.72	11.580	12.660	13.6200	16.49	2.0400	12.59	12.56	3	1.62	7.77
50-64	142	8.55	11.640	12.885	13.7350	17.14	2.0950	12.75	12.73	3	1.69	8.59
64+	68	9.42	11.765	13.215	14.4725	16.03	2.7075	13.08	13.15	3	1.61	6.61

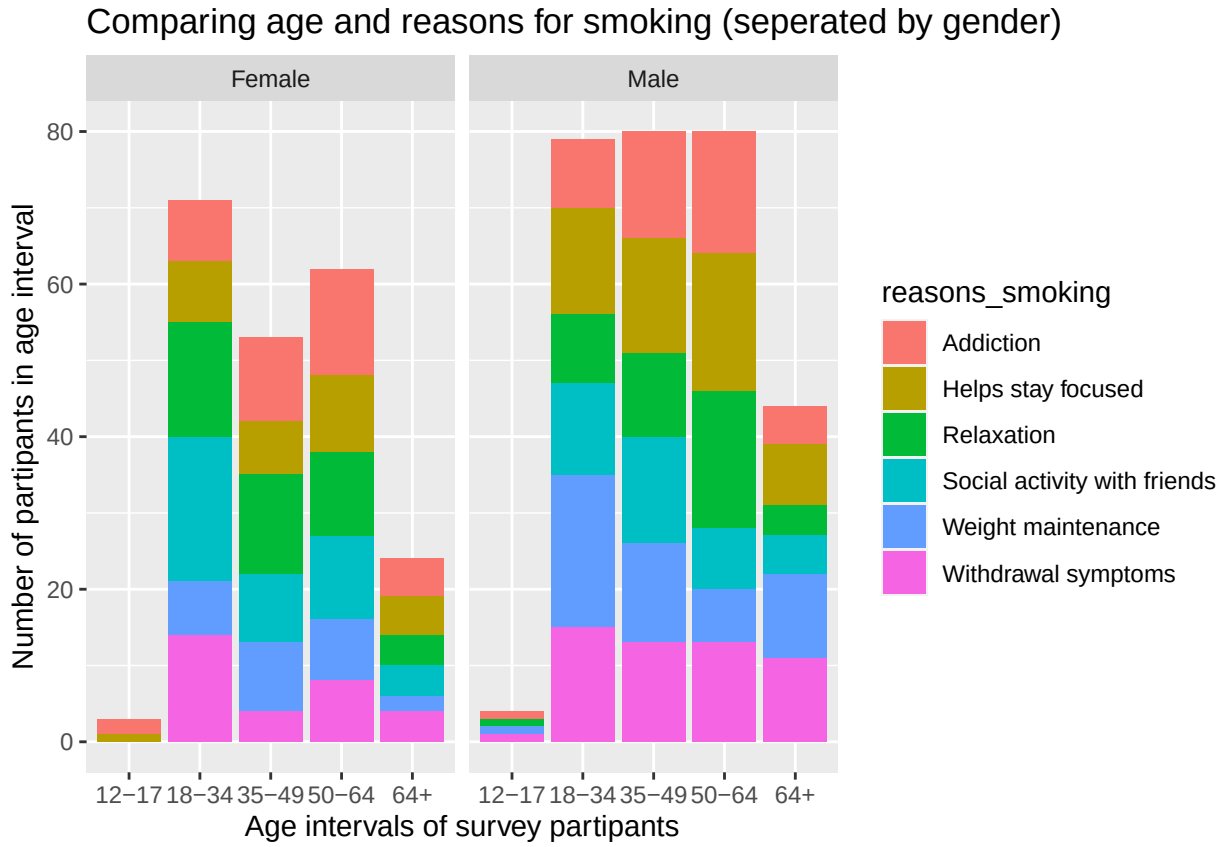
From table_2 we can analyze the numerical summaries of the average number of cigarettes smoked in one day compared to the age intervals of the simulated survey participants. From count we see that most of the simulated survey participants are between the ages 18 to 34 and the least amount of participants are between the ages 12 to 17. This makes sense as under Canadian law it is illegal to sell cigarettes to anyone under the age of 19. The bulk of the simulated survey participants are between the ages of 18 and 64. From table_2 we see that those over 65 have the highest mean, trimmed mean and IQR values of the number of average number of cigarettes smoked in one day, while those between 12 and 17 have the lowest aforementioned values. We see this trend continue for the values between q1 and q3 for all age intervals. However, those between 12 to 17 had the highest minimum value and those between the ages of 50 to 64 had the highest maximum values. This is influenced by the fact in the number of simulated survey participants between the ages of 12 to 17 and those above the age of 65 is much lower than the number of participants in the other age intervals. This leads to the participants between 12 and 17 having a lower value of variability and standard deviation compared to the other intervals.

Graphical Summaries

In addition to numerical summaries we can also use graphical summaries to understand the spread and location of our variables in a visual manner.

(Graph_1)

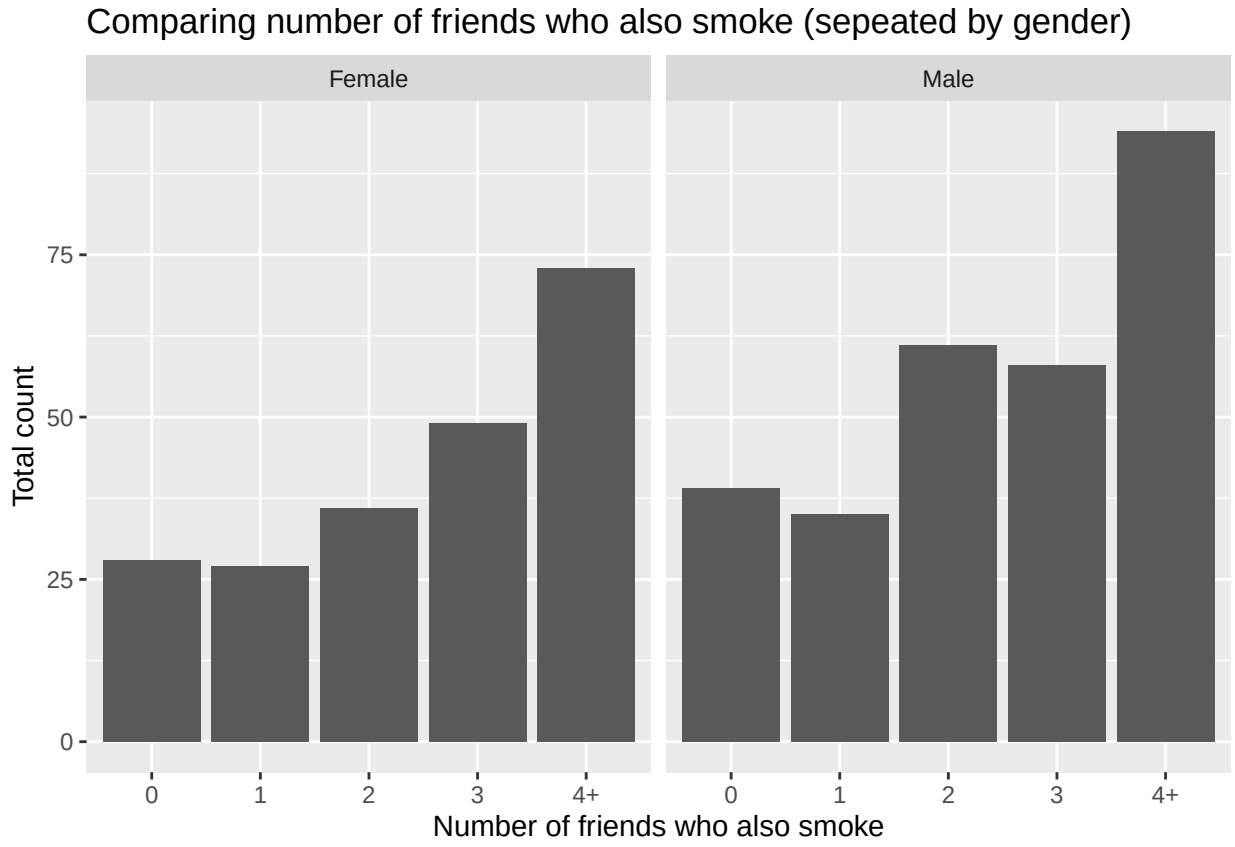
Comparing age and gender to the most significant reason as to to why the simulated survey participants continue to smoke



In graph_1, we see the total number of people in the dataset in each age interval and their most significant reasons for continuing to smoke, separated by gender. From the graph we see that for males, there is roughly an even number of people between the ages of 18 and 64 with roughly half as many people over the age of 64 compared to the total number of people in each of the three age intervals. In addition, we see that the number of males between the ages of 12 and 17 is much lower than all other age intervals. In contrast, we see that most females in the survey are between 18 and 34 years old, with the number of females between the ages of 50 and 64 having the second largest proportion of sample population and those between the ages of 25 and 49 having the third largest proportion of sample population. Similarly compared to the males, we see that those between the ages of 12 to 17 and those above the age of 64 represent much lower proportion of the sample population compared to the other age intervals. As for the most significant reasons as to why they continue to smoke we see that for both males and females, that all of the six reasons are roughly equally proportioned for all age intervals.

Comparing total count of number of friends who also smoke separated by gender

(Graph_2)



In graph_2 we analyze the number of friends each simulated survey participant has that also smokes cigarettes separated by gender. As for the female survey participants, we see a positive linear association trend comparing the total count of the number of friends who also smoke cigarettes and an increasing number of friends who also smoke. From the graph we see that having four or more friends who also smoke has the greatest total count while having zero or one friends who also smoke has the lowest total count in the simulated data. The positive linear association is also seen with the male participants with four or more friends who also smoke having the highest total count in the data while having zero or one friend who also smoke having the lowest total count. This makes reasonable sense as to many people smoking is a social activity done with friends, so it makes sense that smokers would more likely be friends with other smokers.

Concludions for the data section

To conclude the data section, after simulating the data we calculated various numerical summaries of the data to better understand the data. We compared gender to the average number of cigarettes smoked in one day and found that the survey contains more males than females and that on average males smoked more cigarettes compared to females. In addition, we calculated the numerical summaries comparing age intervals of the participants by the average number of cigarettes they smoked in a day. From this we found that on average those between the ages 18 and 34 smoked the highest number of cigarettes in one day. Moving on, we then analyzed a graph comparing age, gender and most significant reason for smoking. We found that for both males and females, that for every age interval the most significant reason for continuing to smoke was evenly proportioned between all six options given in the survey. Finally, we analyzed a graph comparing the the total count of the number of friends who also smoke in the data separated by gender. We found that both males and females had a positive linear association between total count and the number of friends who

also smoke.

All analysis for this report was programmed using `R version 4.0.2`.

Methods

After cleaning and analyzing the data we are able to use frequentist methodologies to analyze and compare the average number of cigarettes smoked in a day for all collective males and females in our simulated data.

Frequentist model: Comparing average number of cigarettes smoked for all genders

Using a frequentist Poisson model we are able to estimate the collective average number of cigarettes smoked in a day for all genders in the simulated dataset and compare the values. To find these values we need to use the following model:

$$Y_{g,c} \sim Pois(\lambda_c)$$

Within this Poisson model, $Y_{g,c}$ denotes a random variable that represents the average number of cigarettes smoked in a day by an individual who belong to gender g and $Pois(\lambda_c)$ denotes the collective average number of cigarettes smoked in a day by all individuals of gender g from the simulated data.

Parameter of interest

Within our Poisson model, our parameter of interest is $\hat{\lambda}_c$. $\hat{\lambda}_c$ represents the maximum likelihood estimate of our model. Using the simulated data and applying the maximum likelihood estimation procedure we are able to find the value of the parameter of interest which represents the collective average number of cigarettes smoked in one day for those who belong to gender g. The calculation for the maximum likelihood estimation procedure can be found in the appendix.

Methodologies

In our dataset using simulation we were able to simulate the average number of cigarettes smoked in a day and gender for all 500 survey participants ($Y_{g,c}$). Using our simulated data we are able to use the maximum likelihood estimator method to find the value of our parameter of interest $\hat{\lambda}_c$ for each gender. This value represents the collective average number of cigarettes smoked in a day of all survey participants of gender g. It is important to note that since we are using a frequentist modeling approach the parameter of interest will be a numerical constant. The calculation of how the maximum likelihood estimator was derived can be found in the appendix section at the very end.

For our model it is appropriate to use the Poisson distribution since we are measuring a finite number of events in a time interval. In this case we are measuring the average number of cigarettes smoked (finite number of events) in one day (time interval). Lastly, it is appropriate to use the Poisson distribution since the events we are measuring are discrete.

Assumptions

Due to the fact that we are using a frequentist Poisson distribution for our model there are assumptions that we need to be aware of. The assumptions in our model are the following:

- Events that occur in different time intervals are independent from one another.
- Two or more number of events can't occur simultaneously.
- The Expected number of events in each interval of is constant.

Aggregation of data

Table 3: Table contains only first 6 of over 200 rows

Males	Females
12.99	13.81
13.28	13.56
14.30	9.92
13.00	10.51
10.34	10.61
13.77	12.82

From table_3 we are able to see the average number of cigarettes smoked in a day for the simulated participants of the survey separated by gender. To calculate the collective average number of cigarettes smoked in day the study will using all 500 simulated participants, however please note that the data displayed in table_3 only displays the first six values for each gender. The values in table_3 are meant to be a glimpse of all values used in the calculation and do not represent all the values used in the calculations. We will be using all the values in our dataset to calculate the collective average number of cigarettes smoked in a day for each gender. From the numerical summaries section we have observed that the females has 213 observations and males have 287 observations. We will be using the maximum likelihood estimator to find the collective average number of cigarettes smoked for each gender. The calculation of how the maximum likelihood estimator was derived can be found in the appendix section at the very end.

Confidence Intervals

After finding the collective average number of cigarettes smoked in one day for each gender we will utilize the Z/t approach to to derive a 95 percent confidence interval (CI) for the calculated mean values. The Z/t approach will involve using the equation: $\hat{\lambda}_c \pm 1.96 * \sqrt{\hat{\lambda}_c/n}$ [9].

Hypothesis Test

From our results we are able to calculate the collective average number of cigarettes smoked in a day for each gender. Using two sided hypothesis test we are able to test our hypothesis on if there is a significant fluctuation between the collective average number of cigarettes smoked between the males are females of the simulated study.

In the hypothesis test our null hypothesis and our alternative hypothesis are the following:

Null hypothesis (H_0): $\mu_f = \mu_m$

In the case of the null hypothesis, μ_f is the true population collective average of the number of cigarettes smoked in a day by females and μ_m is the true population collective average of the number of cigarettes smoked in a day by males. This hypothesis states that there is no significant deviations between the collective averages of both genders.

Alternative hypothesis (H_A): $\mu_f \neq \mu_m$

In the case of the alternative hypothesis, μ_f is the true population collective average of the number of cigarettes smoked in a day by females and μ_m is the true population collective average of the number of cigarettes smoked in a day by males. This hypothesis states that there is a significant deviations between the collective averages of both genders.

Concluding methods section

To conclude, after cleaning the data and preforming numerical and graphical summaries we can use models to analyze the smoking habits of the survey participants. In order to do this we will be estimating the collective average number of cigarettes smoked in a day by the males and females of the dataset. We will be using Frequentest Poisson modeling and incorporate a 95 percent confidence interval in addition to our estimated

value. To estimate the value we will be using a maximum likelihood estimation technique, the calculations of our maximum likelihood model can be found in the appendix. Finally, we will be using hypothesis test to test our hypothesis on if there exists any significant fluctuation between the collective average number of cigarettes smoked between the males and females in the study.

Results

Using the data and methods from the previous sections we are able to find the collective average number of cigarettes smoked in a day by each gender. In addition we are able to calculate a 95 confidence interval for our estimated value and by using hypothesis testing identify if there is a significant change between the average number of cigarettes smoked in one day between males and females.

Table 4: Collective average number cigarettes smoked in one day for all genders

Gender	daily_avg	Confidence_interval_95_percent
Males	12.84	(12.43,13.25)
Females	12.69	(12.21,13.17)

From table_4 we are able to see the collective average number of cigarettes smoked in a day for males and females incorporating all 500 simulated survey participants. In addition, to there is also included a 95 percent confidence interval for the estimated values. With the confidence intervals we can say that we are 95 percent confident that for each gender the interval given will include true population parameter λ_c.

Interpreting of the results From table_4 we see collectively the males in the simulations smoke an higher average amount of cigarettes in one day compared to the females in the survey. These values also correlate with what we have seen in the numerical analysis section in table_1, where we also saw that the mean of the average number of cigarettes smoked in one day by the males is higher than the females. Our take away from this information is that on average Canadian males tend to smoke more cigarettes compared to Canadian females. These values make intuitive sense if we compare the values to a 2020 Statistics Canada survey where they found that there were more males who smoked cigarettes compared to females [2]. The results found in the study match with the results of our simulated survey.

Hypothesis Test Results

From our hypothesis test, using the t.test function in R we have found the p value to be 0.3108 when comparing the collective average number of cigarettes smoked by males and females. Since the P value is small, we reject the null hypothesis and accept the alternative hypothesis. From these results we can surmise that there is a significant fluctuation between the collective average number of cigarettes smoked in a day by males compared to females. This result makes intuitive sense when comparing the values to a 2016 Statistics Canada health survey which found that 23.5 percent of males smokers were heavy smokers compared to 14.2 percent of female smokers who were considered heavy smokers [10]. The finding of the survey conducted by Statistics Canada support the hypothesis test's findings that in Canada, males on average smoke a significantly amount more amount of cigarettes more than Canadian females.

Concluding the results- main takeaways

From our survey results our estimated values, we have seen that males on average smoke more cigarettes in one day compared to women. In addition, with our hypothesis we have found that with a p-value of 0.3108 that there is a significant fluctuation between the average number of cigarettes smoked between the males and females. From these results we can conclude that on from the simulated results of the survey that males on average smoke more cigarettes compared to females.

Bibliography

1. Reid JL, Hammond D, Tariq U, Burkhalter R, Rynard VL, Douglas O. (2019) *Tobacco Use in Canada: Patterns and Trends*. 2019 Edition. Waterloo, ON: Propel Centre for Population Health Impact, University of Waterloo. 2019. https://uwaterloo.ca/tobacco-use-canada/. (Last Accessed: Oct 1, 2021)
2. (2021, Sept 8) *Smokers, by age group*. Statistics Canada. Table 13-10-0096-10 Smokers, by age group. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1310009610. (Last Accessed: Oct 1, 2021)
3. *Smokers helpline*. Canadian Cancer Society https://www.smokershelpline.ca/forums. (Last Accessed: Oct 1, 2021)
4. *Cigarettes: the hard truth*. Canadian Cancer Society. https://cancer.ca/en/cancer-information/reduce-your-risk/live-smoke-free/cigarettes-the-hard-truth. (Last Accessed: Oct 1, 2021)
5. (2013, Nov) *Quitting Smoking in Ontario*. Ontario Tobacco Research Unit. https://www.otru.org/documents/quitting-smoking-in-ontario/. (Last Accessed: Oct 1, 2021)
6. (2020, July 20) *Sex at birth and gender: Technical report on changes for the 2021 Census*. Statistics Canada. https://www12.statcan.gc.ca/census-recensement/2021/ref/98-20-0002/982000022020002-eng.cfm. (Last Accessed: Oct 1, 2021)
7. Hitchman SC, Fong GT, Zanna MP, Thrasher JF, Laux FL. (2014, May 19) *The Relation Between Number of Smoking Friends, and Quit Intentions, Attempts, and Success: Findings from the International Tobacco Control (ITC) Four Country Survey*. Psychol Addict Behav. 2014;28(4):1144-1152. doi: 10.1037/a0036483. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4266625/. (Last Accessed: Oct 1, 2021)
8. Schenkelberg F. *The Range Rule*. Accendoreliability. https://accendoreliability.com/the-range-rule/. (Last Accessed: Oct 1, 2021)
9. Dr Deng. (March 20 2014). *Computing confidence interval for poisson mean. Computing Confidence Interval for Poisson Mean*. http://onbiostatistics.blogspot.com/2014/03/computing-confidence-interval-for.html. (Last Accessed: Oct 1, 2021)
10. Janz T. (2015, Nov 27) *Current smoking trends*. Statistics Canada Catalogue no. 82-624-X. https://www150.statcan.gc.ca/n1/pub/82-624-x/2012001/article/11676-eng.htm. (Last Accessed: Oct 1, 2021)

Appendix

Here is a glimpse of the data set simulated/surveyed:

```
## Rows: 500
## Columns: 7
## $ age                 <fct> 35-49, 35-49, 50-64, 50-64, 50-64, 18-34, 50-64, 5~
## $ gender              <fct> Male, Male, Male, Male, Male, Male, Male, Female, ~
## $ quitting_intentions <fct> No, Yes, Yes, Yes, Yes, No, No, Yes, No, Yes, Yes,~
## $ reasons_smoking     <fct> Addiction, Withdrawal symptoms, Withdrawal symptom~
## $ average_daily       <dbl> 12.99, 13.28, 14.30, 13.00, 10.34, 13.77, 12.78, 1~
## $ friends_smoke_count <fct> 2, 4+, 2, 0, 4+, 2, 0, 3, 4+, 1, 2, 2, 4+, 4+, 2, ~
## $ ecig_use            <fct> Yes, No, No, Yes, Yes, Yes, No, No, No, Yes, Yes, ~
```

Calculating Maximum likelihood estimator for Poisson distribution.

n = Total number of observations

$P(X_i = x_i|\lambda) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$

finding the likelihood of the function.

$L(\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$

$L(\lambda) = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x!}$

Taking the log of the function.

l(λ) = -nλ + ($\sum_{i=1}^{n} x_i$)log(λ) - $\sum_{i=1}^{n} x_i$!

Taking the derivative

l'(λ) = -n + ($\sum_{i=1}^{n} x_i/\lambda$)

-n + ($\sum_{i=1}^{n} x_i/\lambda$) = 0

Finding the estimated value function

$\hat{\lambda} = \sum_{i=1}^{n} x_i$/n

We need to use the second derivative test to verify its the maximum.

l"(λ) = -$\sum_{i=1}^{n} x_i/\lambda^2 < 0$

Since all values of x_i are greater or equal to zero and due to the fact that the square assures us λ^2 will always be positive we can conclude that

l"(λ) = -$\sum_{i=1}^{n} x_i/\lambda^2$ will always be less than 0 and therefore we can conclude that the value of MLE for the Poisson distribution is $\hat{\lambda} = \sum_{i=1}^{n} x_i$/n