STA304 - Fall 2021

Assignment 1

[Erica Zhou - 1005687678]

Part 1

Goal

Nowadays, more and more chat apps come into our lives. Especially during the pandemic, there is a significant increase in the demand for chat apps. For popular chat apps, we want to find which is the most popular. We are also interested in the reason why people prefer one app to the others. The goal of the survey 'Chat App Preference' is to collect information on people's preferences for some popular chat apps. The survey can give us an overview of how different genders and age groups prefer different chat apps and their reasons. The survey aims to provide data for analyzing the relationships between ages and time spending on chat apps, ages and preferences, and so on. With this information, we can find out the effect of different factors on the preferences for the apps, and thus we can give the information to the developers and help the improvement and upgrades of the apps.

Procedure

The survey is created for all people aged above 10 in Canada. Ideally, the survey should cover a big number of participants of all age groups above 10. The frame and sample population should at least cover all age groups to avoid selection bias. However, due to the limits of time and budgets, I decided to simulate the data. The simulation is completely random with a size of 500, thus, the observations would be more reasonable because a larger simulation size will cover more random observations. The advantage of simulating data rather than collecting from the survey is that we can avoid selection bias and nonresponse bias. However, the drawback is the lack of accuracy. We cannot ensure that the data simulated is ever close to the actual ones. To give the data a bit more reality, I form my simulation partially based a popularity ranking of chat apps in Canada (will be described in Part2).

Showcasing the survey.

Survey: Chat App Preference

https://wzngpedipag.typeform.com/to/Q8fK1Gxr [1].

Question 1: How much time do you spend on chat apps every day?

It is a general question provided for the participants to get into the survey. However, it can also give us information about how much people rely on chat apps. With the information, the developers of the chat apps can make improvements, such as adding an eye protection mode to the apps. For this question, a participant may choose a number (integer) between 0 and 5 representing the hours spending on chat apps every day. 0 means 'do not use chat apps' and 5 means '5 or more hrs every day'.

The drawback of this question is that it doesn't give us detailed information about a specific app like 'the time spent on Discord'. However, I think the drawback is acceptable because the survey doesn't focus on the time information, and this question acts more as a transition to some more subjective questions after it. I think it is a reasonable question to put at the beginning of the survey.

Question 2: Which is your favorite chat app?

It is an important question we are interested in. It is a multiple-choice question with a single selection. A participant can choose one from WhatsApp, Facebook Messenger, WeChat, Line, Discord, Skype, and Other to be their favorite chat app. The list includes 6 popular chat apps in Canada and an 'Other' option. It is not a time-taking question, but it can provide us with the most direct information about the preference.

The drawback of this question is that the list doesn't cover all the popular chat apps. And since it is only available for a single selection, some participants may have trouble when deciding the favorite one.

Question 3: Why is it your favorite?

It is the question after Question 2 as a complementary question that digs deeper into the preference of a participant. It is also a multiple-choice question with a single selection. Participants may select one from 'Easy to use, Widely used by my family and friends, Nice security, Beautiful layouts, Unique functions (e.g. memes, translation. . .), and Other'.

This question can provide data to support the development of chat apps. If a lot of people prefer the app because of its beautiful layouts, then the developers of chat apps can focus on the improvement of the layouts of the app. The drawback of this question is the single selection. Participants may have trouble selecting only one option.

Part 2

Data

The variables are collected from the survey 'Chat App Preference'. The data simulated is completely random combined with numerical and categorical variables. However, I simulated 'fav_apps' with different probabilities of the apps based on the app ranking on https://www.similarweb.com/apps/top/google/store-rank/ca/communication/top-free/ [2]. Therefore, most of the observations are not reliable. However, this project provides a frame on this topic for any further work and reproduction. Ideally, the data should be collected directly from the survey.

Assumed probabilities:

Option	Probability
WhatsApp	0.35
Facebook Messenger	0.15
WeChat	0.05
Line	0.05
Discord	0.25
Skype	0.10
Other	0.05

I converted the values to the integers that rank the age groups from 1 to 9 the in the column of age_group so that the age groups become continuous.

Important variables

age_group: It is a numerical variable that represents 5-year age groups above 10 of the participants. The age groups are represented by levels from 1 to 9. For instance, level 1 means '10 to 15(inclusive)', level 2 means '16 to 20', and level 9 means 'above 50(exclusive)'.

hours: It is a numerical variable that represents the average hours spent on chat apps every day by the participants. It is ranged from 0 to 5 (5 includes 5 or more hours).

gender: It is a categorical variable that represents the gender of the participant. Participants can choose 'Male', 'Female', 'Non-binary' or 'Other(Not listed above)'.

num_of_apps: It is a numerical variable that represents the number of chat apps that a participant has on his/her smart devices. Participants can choose from 0 to 5 (5 includes 5 or more apps).

fav_apps: It is a categorical variable that represents the favorite chat app chose by a participant. Participants may choose from 'WhatsApp, Facebook Messenger, WeChat, Line, Discord, Skype, and Other'.

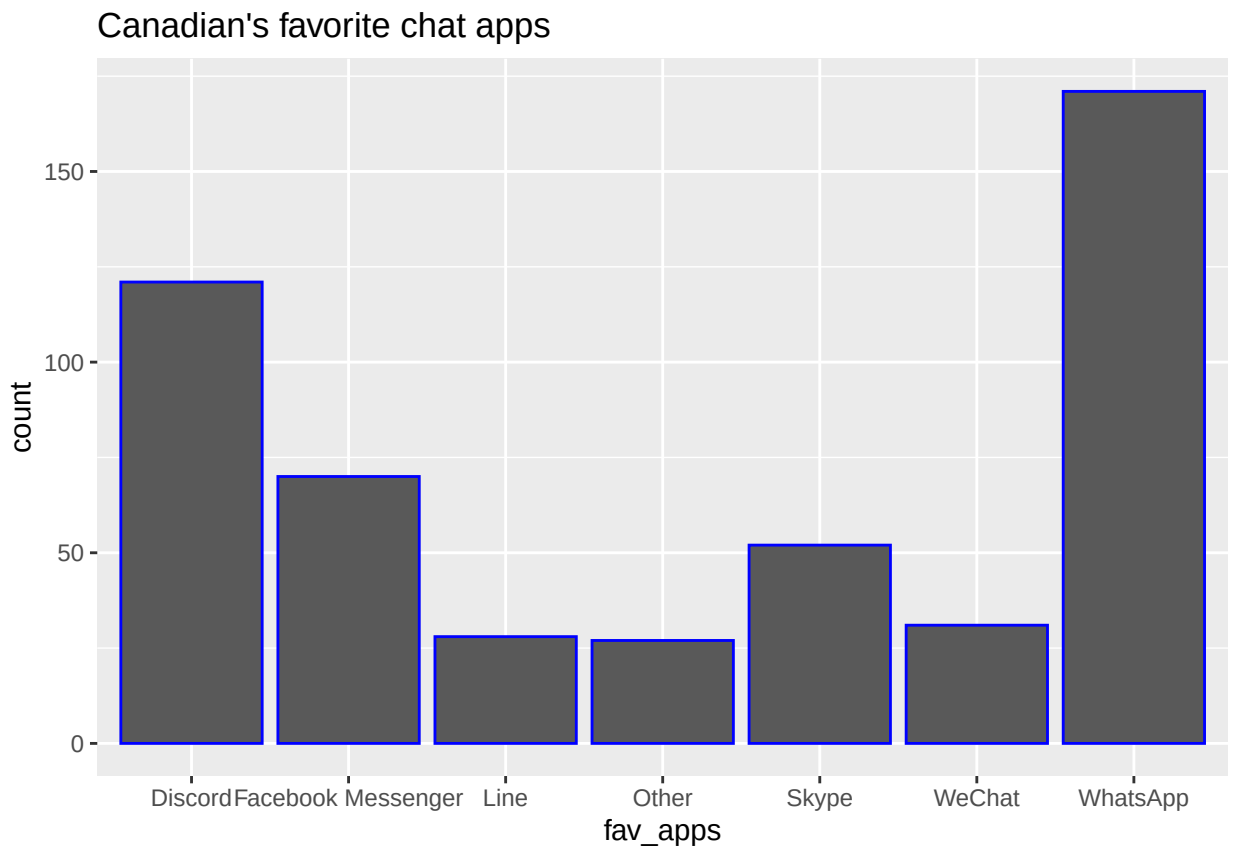
reason: It is a categorical variable that represents the reason why a participant selects one chat app to be the favorite. Participants may choose from 'Easy (to use), Widely used (by my family and friends), Nice security, Beautiful layouts, Unique functions, and Other'

suggestion: It is a categorical variable representing the suggestion a participant may have towards his/her favorite chat app. Participants may choose from 'Bug reduction, Better information protection, Better layouts, More functions, No suggestion (It's already perfect!), and Other'.

Variable	Mean	Standard Deviation
hours	2.976	1.431
num_of_apps	3.056	1.427
age_group	4.966	2.542

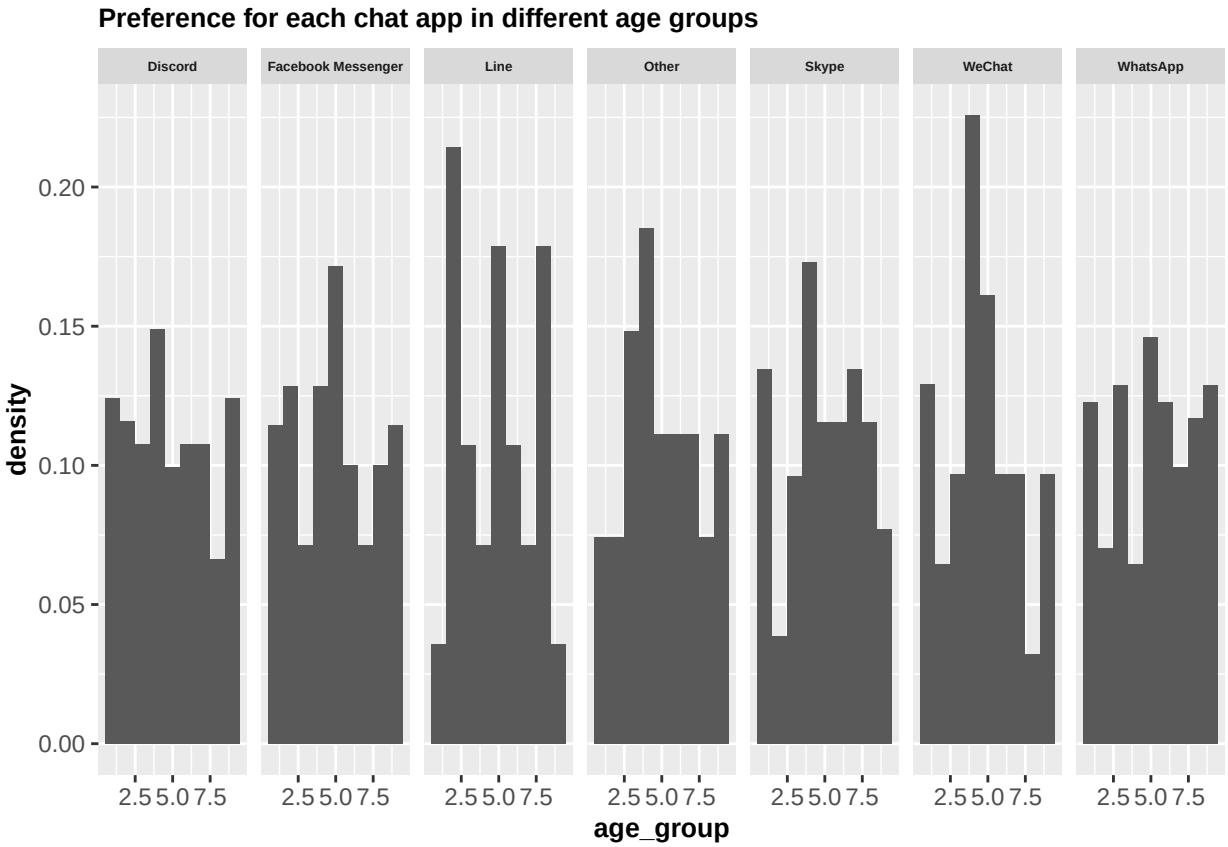
(The table only bases on the simulation data.)

```
df %>% ggplot(aes(x = fav_apps)) + geom_bar(col = 'blue') + ggtitle("Canadian's favorite chat apps")
```



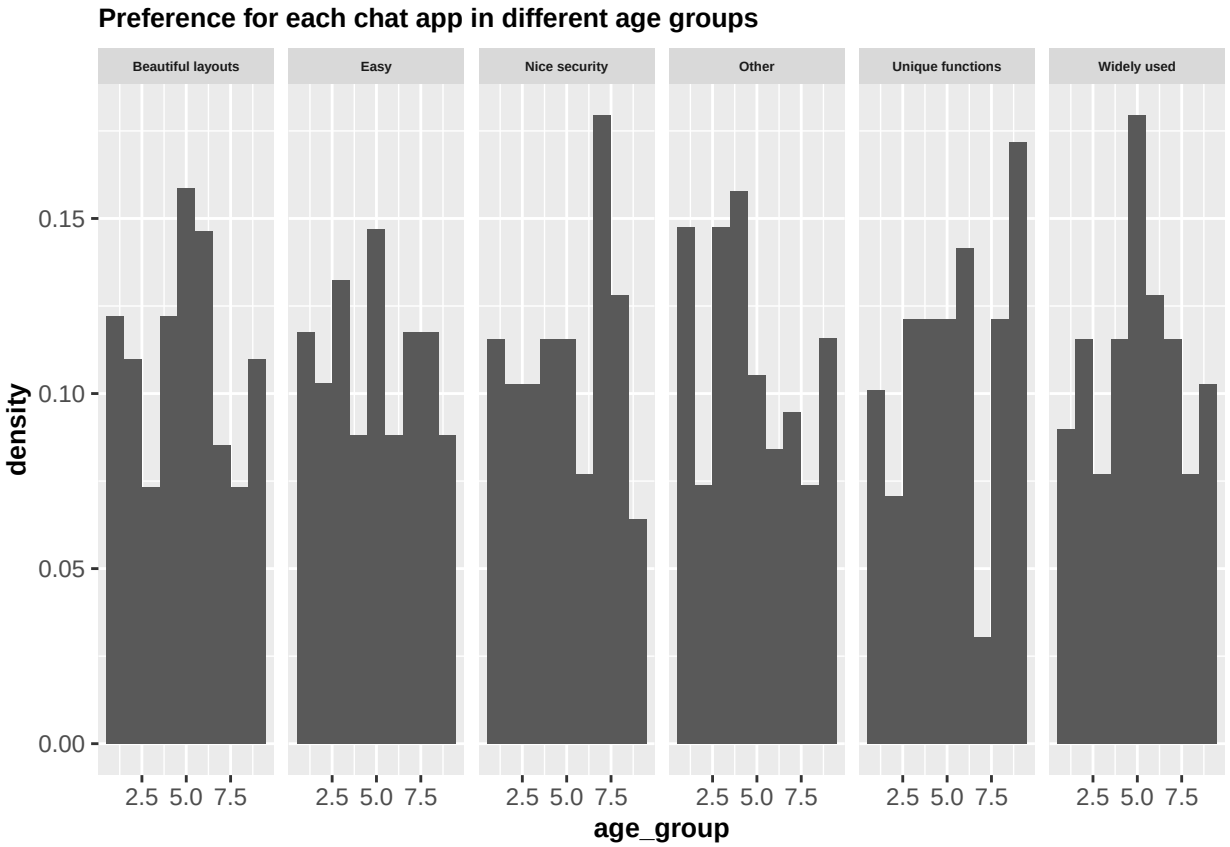
The barplot represents the number of people that choose each app to be their favorite. We see that Whatsapp is the most popular chat app in Canada. Discord and Facebook Messenger are also more popular than others.] Line and Wechat are less popular because they are most popular among Japanese or Chinese people but not that popular among Canadians. The outcome seems reasonable because I simulated the data based on the existing popularity ranking mentioned before.

```
df %>% ggplot(aes(x = age_group, y =..density..)) + geom_histogram(bins = 9) +
  facet_grid(col = vars(fav_apps), scales = 'free_y') +
  ggtitle('Preference for each chat app in different age groups') +
  theme(strip.text.x = element_text(size = 5, face = "bold"), axis.title=element_text(size=10,face="bol
        plot.title = element_text(size = 10, face = "bold"))
```



The histograms show how people in different age groups prefer a chat app. The outcomes don't seem entirely reasonable if we look at the outcomes of 'Skype', a big portion of the youth aged between 10 to 15 prefer Skype to others. As we know, Skype is an early video chat app released in August 2003. Normally speaking, most Canadian teenagers are not likely to prefer Skype. Because I simulated the data randomly, the outcome is unreal. The outcome of 'WhatsApp' seems more reasonable because it could be equally popular among all age groups in practice.

```
df %>% ggplot(aes(x = age_group, y =..density..)) + geom_histogram(bins = 9) +
  facet_grid(col = vars(reason), scales = 'free_y') +
  ggtitle('Preference for each chat app in different age groups') +
  theme(strip.text.x = element_text(size = 5, face = "bold"), axis.title=element_text(size=10,face="bol
        plot.title = element_text(size = 10, face = "bold"))
```



The histograms show why people select the app to be the favorite. Under the category 'Nice security', we see that a big portion of the mid-age groups and the groups above 50 appreciate his/her favorite chat app because of its security. It seems reasonable in practice because when ages increase, people have more personal information needed to be protected such as wages and properties. Nevertheless, the outcome of the category 'Easy'(to use) doesn't seem reasonable because most elderly people need more convenient functions than young people when using a chat app. Thus, there should be a higher bar on age group 9 in practice in this category.

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

Methods

I am going to focus on the most popular chat app 'WhatsApp' and find out if it 'wins' the game because of its unique functions. I am going to apply a hypothesis test and confidence interval to see how the population perform.

A hypothesis test [3] is a statistical testing that can determine whether there is a significant level of evidence to support or reject the hypothesis under the study. A confidence interval [3] is an interval in which we are confident that the population parameter locates.

Hypothesis test:

Let p be the probability that a Canadian like WhatsApp best among all popular chat apps because of its unique functions.

Our null hypothesis is that $p = 1/5$ (assumption)

H_0: $p = 1/5$

Our alternative hypothesis is that $p < 1/5$

H_A: $p < 1/5$

We are interested in if there is significant evidence to support or reject the null hypothesis. In this case, we apply a hypothesis test to see if the population probability is smaller than 1/5.

Since there are 5 reasons ('Other excluded') mentioned in the survey, I assumed that the probability that an individual appreciate the unique function of WhatsApp to be 1/5 (p_0). Then I need to know the sample statistics. The number of observations is the number of participants, who prefer WhatsApp (n), and the sample mean is calculated through dividing the number of participants, who like the unique function of WhatsApp by n. Because there are only 2 cases according to our assumption, 'because of unique functions' or 'not because of unique functions'. $\hat{p}$ is the estimate calculated through sample mean. The sample variance is then $p_0(1-p_0)$. After that, plug the values in to the test statistic and use the Z-score test[3]:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0,1)$$

and the p-value is calculated by P(|Z| < z) if z is negative.

Confidence interval:

I will invoke a non-parametric bootstrap [2] to derive the 95% confidence interval (CI) for the probability that a Canadian like WhatsApp because of its unique functions.

The 95% CI is

$$CI : [\bar{x_n} - c_u^* s_n/\sqrt{n},\ \bar{x_n} - c_l^* s_n/\sqrt{n}]$$

where c_l^* and c_u^* are estimated by the 0.025 and 0.975 percentiles of the bootstrapped studentize means. $\bar{x_n}$ is the bootstrap sample mean, s_n is the standard deviation, and n is the number of observations.

Results

Concept	Result	Interpretation
95% CI	[0.111, 0.218]	We have 95% confidence that the true probability lies between 0.111 and 0.218.
p-value	0.118	The p-value is larger than the 0.05 significance level, thus there is a weak evidence against the null hypothesis, that is, a weak evidence that the probability is smaller than 1/5.

The 95% CI tells us the approximate range of the parameter p. However, the p-value test is not very effective based on this simulation. In further studies, researchers should make more suitable assumptions about the parameters and use real data from the survey. In conclusion, this project gives an overview of the preference of Canadian people for chat apps. WhatsApp is the most popular chat apps in Canada. Nevertheless, the reason why people prefer WhatsApp still needs to be explored in the future with actual data. In future studies, researchers can also focus on the popular reason not limited to any specific app but all the chat apps, so that we can know in what direction the chat apps should be improved.

Bibliography

1. Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: May 5, 2021)
2. Similarweb (2021, September 28). *Top communication apps ranking - most popular apps in Canada.* Similarweb. https://www.similarweb.com/apps/top/google/store-rank/ca/communication/top-free/.(Last Accessed: Sep 30, 2021)
3. Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how.* Springer Science & Business Media.

4. Allaire, J.J., et. el. *References: Introduction to R Markdown*. RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: May 5, 2021)

Appendix

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

```
## Rows: 500
## Columns: 7
## $ age_group   <dbl> 9, 4, 7, 1, 2, 7, 2, 3, 1, 5, 5, 6, 7, 9, 5, 5, 9, 9, 5, 5~
## $ hours       <int> 2, 1, 3, 1, 3, 5, 2, 4, 1, 4, 1, 1, 4, 4, 5, 3, 2, 2, 3, 1~
## $ gender      <fct> Female, Male, Male, Non-binary, Female, Non-binary, Non-bi~
## $ num_of_apps <int> 3, 5, 2, 4, 1, 1, 3, 1, 2, 5, 1, 3, 1, 3, 3, 3, 5, 5, 4, 1~
## $ fav_apps    <fct> WhatsApp, Facebook Messenger, WhatsApp, Facebook Messenger~
## $ reason      <fct> Beautiful layouts, Unique functions, Nice security, Nice s~
## $ suggestion  <fct> Bug reduction, Better layouts, More functions, Other, Bett~
```