STA304 - Fall 2021

Assignment 1

Yutong Lu 1005738356

Part 1

Goal

This survey aims to investigate how the COVID-19 global pandemic has impacted the spending behaviours on mobile games for mobile game players in Canada. This is particularly interesting because the NPD group (2021) reports that in the second quarter of 2021, one year after the pandemic began, there still appears to be a continued increase in mobile game spending in the U.S. [14]. Simultaneously, according to the analysis by International Data Corporation and LoopMe in 2021, they reported a substantial increase not only in the total number of mobile game players worldwide but in the proportion of players increasing their gameplay time since the pandemic began [16].

This survey focuses on the financial aspects of mobile gaming, with questions concerning the income levels and monthly budgets of mobile gamers. More importantly, it asks about the changes in the monthly spending on in-game purchases and weekly gameplay time before and after the pandemic happened. The familiarity with profitable online gaming-related platforms such as live stream and gaming channels that has also increased in popularity since COVID-19 began, is also investigated in this survey. Notice that this survey is only concerned with the spending on in-game purchases and not with payments made to download the games, so it means the money spent on the games themselves is not a part of the survey.

By analyzing the responses for these questions, we may get a better idea about the different factors brought by the pandemic that can affect people's spending habits and their extent of impact on people, which contributes to the overall topic about how the COVID-19 has changed the way people spend on mobile games.

Procedure

To implement the survey, I will send out target surveys to people on different platforms, including people creating or watching gaming content on video websites such as YouTube, users in gaming-related Reddit forums, viewers and streamers on live streaming platforms. To reach less active participants in the online communities above, I will also directly send the survey in message applications and social media to ask people to help share this survey to other people to reach mobile gamers that may not be reached on gaming-related platforms.

In terms of the population, the target population is all the mobile game players in Canada who have played a mobile game with in-game purchase options. However, there could be individuals entering or leaving Canada as this survey is still ongoing. Also, it is difficult to define the term "mobile game players" since there may be individuals who used to play such mobile games but stopped playing before the global pandemic. Therefore, the frame population is all the current active players in Canada who are playing mobile games with in-game purchase options. On the other hand, because the survey is mainly implemented through emails or other online platforms related to gaming, we may not reach some parts of the population. For example, using our procedure, we are less likely to reach the younger individuals whose usage of electronic products and

in-game spending are under the guidance of their parents, players who do not actively engage in such online communities, or players who spend on casual games like Candy Crush Saga that do not necessarily have such a community on, say, Reddit. As a result, the sample population is all the active players in Canada who play mobile games with in-game purchase options and are participants of online activities to some extent such that this survey can reach them.

The strength of the proposed sampling procedure is that due to the spreadability of online information, it is relatively easy to reach a significant number of individuals in a short time using online distribution in communities, group chats, and forums. Also, because I not only send the survey to online communities but personal message applications and social media, I can still possibly reach people who are not active participants of gaming communities and get responses from people in different age groups and fields. Also, the sampling is meant to restrict to only people in Canada. This is because we want to reduce the differences in the responses due to mobile gaming market differences around the world. Also, because there will be a question in the survey asking about the COVID-19 severity in the observation's area, it could be very difficult to evaluate pandemic severity using a simple scale in different parts of the world and account for its impact on mobile game spending. Another reason for restricting the sample frame to mobile game players in Canada is that if the survey is conducted at a worldwide level, then there may be language barriers for mobile game players that are non-English speakers and thus we may fail to gather their responses and eventually miss lots of information because they also make up a significant part of the mobile gaming community.

The drawbacks of this procedure are mainly caused by the online nature of this survey. It could be a form of non-probability sampling, more specifically, restricted and volunteer-based sampling because we may miss individuals who simply play and spend in mobile games but do not actively use message applications or engage in online communities that are gaming-related. As a result, gamers below and above certain ages may not be reached by this survey, but they also may be spending a substantial amount of money on mobile games. Simultaneously, the participation in online gaming communities may imply a greater devotion to the games, so it may naturally lead to us sampling individuals that are heavier spenders on mobile games and missing individuals that are casual spenders. Another concern is that because the information is distributed online, it cannot guarantee that only people in Canada have filled in the survey even though it is specified on the survey title. Also, because I sent out the survey from my personal social media account first, it could be convenience sampling as this survey will essentially reach my friends and family and people that are associated with them in some ways, and thus only parts of the population are included in the sample.

Showcasing the survey

Survey Link

https://docs.google.com/forms/d/e/1FAIpQLSeWViqOY5-jKN7r_3Fi7P_A7IhvBP3113SfRdHczGljLJ9WEw/viewform[1]

I chose these three questions because they reflect the relationships I want to capture between the global pandemic, the changes or "new normal" brought by it, and the mobile gamers' spending behaviours on in-game purchases.

Question: On average, how much money (in CAD) do you spend per month on in-game purchases in mobile games now, i.e., after the pandemic began?

Response: fill in the blank

The first question is a numerical question in a four-question group where the survey asks the participant to report the average monthly spending and weekly gameplay time before and after the global pandemic began. This question, in particular, is asking about the monthly spending on the in-game purchases of mobile games, and its response will be used to compare with the response for monthly spending before the global pandemic. The benefit of this question is that we can get a direct idea of how much more or less money an individual spends since COVID-19 started by comparing to the response to the question asking about the spending before the pandemic. Also, we can perform hypothesis testing or confidence intervals on the numeric values. However, one drawback of this question is that it does not reflect the actual reason for

the change in spending, and the reason itself may not be directly related to the global pandemic. Also, the self-reported average value may not be accurate due to false information recall or other undisclosed personal reasons.

Question: Do you directly profit from a gaming-related platform to support your in-game purchases (e.g. Twitch live stream, YouTube gaming channels)?

Response: multiple choice

a. Yes, I do, and I started it before the global pandemic.

b. Yes, I do, and I started it after the global pandemic.

c. No, I don't, but I am aware of these platforms, and I watch these streams/channels.

d. No, I don't, but I am aware of these platforms, and I don't watch these streams/channels.

e. No, I don't, and I am not aware of these platforms.

The second question is a categorical question that asks about the participants whether they have profited from any online gaming-related platforms, which may contribute to the spending habits of this individual to some extent. The benefit of this question is related to online platforms that have dramatically increased in popularity since the COVID-19 pandemic, including streaming, gaming channels, and other forms of online platforms. According to Kominers (2021), in only one month between March and April of 2020, there was a 300% increase in the Twitch audience in some categories [5]. As a result, it is possible that some individuals may profit from these platforms via gaming and these revenues, in turn, affect their monthly spending on in-game items of mobile games. Simultaneously, it asks the participants about their awareness and engagement level of these platforms even if they are not creators on such platforms, which allows us to analyze the potential relationship between the engagement of online gaming communities and spending on games. The drawback of this question is that being a live streamer or YouTuber may not guarantee a consistent income that can support their spending on games, especially for smaller creators that just started their channels. Also, the income difference can be huge for these creators. As a result, there still may be a tremendous difference in their monthly gaming budget in the group of people who chose yes for this question.

Question: How severe do you think the COVID-19 pandemic is in your area?

Response: self-report on a scale from 1 to 5, with 1 being not severe at all and 5 being very severe

The third question is a categorical question that asks about the COVID-19 severity in this participant's living area. Ward (2021) reports that mobile gaming activity is affected by the local pandemic severity, and some individuals reported that they engage in games to socialize safely under the social restrictions during COVID-19 [16]. The benefit of this question is that we can investigate the potential relationship between the pandemic severity and spending on in-game purchases by associating the difference in observations' spending before and after the pandemic began (using the first question mentioned above and the other three complementary questions) and the COVID-19 severity of their living areas. The drawback of this question is that the severity of the pandemic, although quantifiable using infected and death numbers, may be subjective to individual feelings and may change during different time periods. The scale of evaluation may be different from one individual to another, where these individuals are also from different areas. As a result, this question may not draw an objective conclusion regarding the severity of the pandemic.

Part 2

Data

Data, Context and Simulation Process

The data is simulated in R [10] with 200 observations and 11 variables in total to represent the potential 200 participants 11 questions in the survey. It contains both numerical and categorical values. The number of observations is chosen based on the potential number of responses I would obtain if I were to distribute this survey during the data collecting period of this report. The simulation techniques and the relevant parameter values involved in simulation are determined based on literature. Because the context of this data is meant to be taken from the sampling frame in Canada, the goal is to gather relevant values for the variables from Canada or the U.S., if possible, when referencing the literature, but data from other countries or worldwide data may be used if the data from North America is not available.

To begin with the data simulation process, first a seed is set to ensure the reproducibility of the results, then a sample size of 200 is defined. The first question of the survey asks about the age of participants, so a vector of factors with levels of different age groups is created, and then 200 age levels are sampled randomly with replacement. The probability of the age groups "0-15", "16-23", "24-35", "36-50", "51-65" and "66 and above" are set to be 0.16, 0.25, 0.35, 0.12, 0.09 and 0.03, respectively. This is based on the gamer age groups in U.S. in 2020, where the age group of "18-24" takes up 24.9% and "25-34" is reported by Andre to take up 35.3% of all gamers [1]. However, "18-24" is the youngest group in the report mentioned above, meaning that there is no data about the age groups below 18 or above 64, so the other percentages are split among the other age groups in my survey [1]. The younger age groups are set to take a higher proportion because the growth of younger mobile gamers is reported to be the most substantial among all age groups in the UK in 2021 [3] based on the report by Clement (2021). The 200 observations are stored in a vector for the variable "age".

The second question is about the participant's gender, and a vector of factors with levels "female", "male", "prefer not to say" and "self-specified" is created. Notice in the actual survey, the "self-specified" category is shown as a textbox and allows the participant to specify. Because we are not able to get the specific genders provided by the participants, every response in the textbox is grouped to "self-specified". Then, 200 observations are randomly sampled with these factors with replacement, and the probability for these categories are 0.39, 0.53, 0.04 and 0.04, respectively. Because gender is considered as binary in most mobile gaming reports found online, the proportion of 0.08 for non-binary individuals is based on Clement's report about the gender distribution of worldwide gaming developers in 2021 [2]. The two categories, "prefer not to say" and "self-specified" are equally probable (0.04 for each) in the simulation because no assumption is considered as appropriate here. The rest of the proportions is calculated using Andre's report that 57.7% male and 42.3% female play mobile games [1]. These 200 observations are stored in a vector for the variable "gender".

The third question is about the income level of the participants, but in the simulation, the actual amount of income is simulated, not the levels. The income levels will be created in the data cleaning process. This is because according to Statistics Canada (2020), the annual average income is 47,300 CAD in 2019 [9]. As a result, a monthly average of 4083.3 CAD is calculated and used as the mean value for a normal distribution. 200 observations are sampled from this normal distribution with a mean of 4083.3 and a standard deviation of 2000, which is stored in a vector for the variable "income".

The question four and five ask about the monthly spending on in-game purchases before and after the COVID-19 pandemic began. For both variables, "money_before" and "money_after", 200 observations are sampled from a normal distribution, respectively. The means of these two random variables are based on the American iPhone users' average monthly in-app spending on mobile games reported by Coulson (2021), which are 53.80 USD (67.94 CAD) in 2019 and 76.80 USD (96.98 CAD) in 2020 [4]. The amount of money is converted into CAD based on the currency rates on September 28, 2021. As a result, 200 observations are simulated from a normal distribution of a mean of 67.94 and a standard deviation of 20 as the monthly

spending on mobile games in 2019, and another 200 observations are sampled from a normal distribution of a mean of 96.98 and a standard deviation of 20.

Question 6 and 7 are concerned with the average gameplay time on mobile games per week before and after the pandemic began. For variables "time_before" and "time_after", 200 observations are simulated for each variable using two normal distributions with different means. As reported by Limelight Network (2019), the average time spent on games is 7.61 hours in 2019 in the U.S. [7]. And in 2021, Limelight Network reports that the average time spent on video games is 7.71 hours in the U.S. [8]. As a result, the mean of the normal distributions for 2019 and 2021 are 7.61 and 7.71, respectively, and a standard deviation of 2 is used in both cases. However, Limelight Network did not specify the types of video game platform of this average time, so this number may include gameplay times from other platforms other than mobile games.

Question 8 asks about whether the participant is profiting from a gaming-related online platform, including livestreaming, gaming channels, and more. If the participant is not profiting from such platforms, then the level of participation and awareness of these platforms are asked. Therefore, a factor vector is created with levels "yes and joined before", "yes and joined after", "no, aware, participate", "no, aware, does not participate" and "no and not aware", with "yes or no" meaning if the participant profits from such platforms, "joined before or joined after" meaning if they joined the platform before or after the pandemic began, "participate or does not participate" meaning if they participates or visits these platforms, and "aware or not aware" meaning whether they know about these platforms. According to TwitchTracker (2021), there is a substantial increase in both Twitch streamers and visitors since 2020 [15]. However, no information is found about the proportion of mobile gamers who participate in these platforms, so no assumption about the probability of each factor is incorporated in the simulation. 200 observations are sampled randomly from the factor vector with replacement and no specified probability to simulate for this question, and they are stored in a vector for the variable "platform".

Question 9, 10 and 11 ask about the severity of COVID-19 in the participant's area, the impact of the new normal on the participant's spending behaviours on mobile games, and the likelihood of continuing spending on mobile games in the future. All three questions are asking the participants to rate a number from 1 to 5 based on their subjective view of the question, with 1 being the lowest extent and 5 being the highest extent. Because these questions are heavily based on personal views and attitudes, no probability is assumed for each level, and thus 200 observations are randomly simulated to be a number from 1 to 5 inclusive with replacement for each question, which are stored in variables "severity_level", "impact_of_new_normal" and "cont_spending".

The Cleaning Process

To clean the data, the R package Tidyverse [17] is used for variable creation and variable selection Firstly, a new variable named "income_cat" is created to categorize the incomes of the 200 observations into income levels "< $1000", "$1000 – 2400", "$2400 – 3000", "$3000 – 5000", "$5000 – 7000" and ">= $7000". This is because in the question 3, where it is asking about the income of the participant, it in fact asks about the income level, not the actual amount. During the cleaning, if the income of the individual falls right on the borderline, then it is going to be categorized into the level that is with the higher income group. For example, an observation with an income of $2400 is categorized into the group "$2400 – 3000". Also, because a normal distribution is used to simulate the income data, there may be negative incomes, and with this cleaning procedure, any negative incomes will be categorized as "< $1000".

Because normal distributions are also used for money and time spent on mobile games, there may be negative values as well, which would be set to 0 in the cleaning process. However, after inspection, no value below 0 is found in the dataset, so no negative value needs to be set to 0. Then, two more variables "money_diff" and "time_diff" are created to represent the difference in the money and time spent on mobile games before and after the pandemic began, which is done simply by using the amount of money or time spent after the pandemic began to minus the amount of money or time spent before the pandemic began.

Finally, because we are only concerned with the income categories, money and time differences in this analysis, only the newly created variables "income_cat", "money_diff", and "time_diff" are selected to be in the final dataset, instead of the original variables of income, time and money generated in the simulation.

As a result, we select 9 variables "age", "gender", "income_cat", "money_diff", "time_diff", "platform", "severity_level", "impact_of_new_normal" and "cont_spending" into the final, cleaned dataset.

Description of Important Variables

The important variables in the dataset are "money_diff", "time_diff", "platform", "severity_level" and "impact_of_new_normal", which relate to the five questions that concern with the change brought by the global pandemic.

The variable "money_diff" is the difference in an observation's monthly spending on mobile games' in-game purchases before and after the pandemic began. It is numeric and can take negative, positive or zero as a value because there may be a decrease, increase or no change in the observation's spending.

The variable "time_diff" is the difference in an observation's weekly gameplay time on mobile games before and after the pandemic began. Similar to "money_diff", it is a numeric variable that can take negative, positive or zero as a value because there may be a decrease, increase or no change in the observation's gameplay time.

The variable "platform" is a factor variable with levels "yes and joined before", "yes and joined after", "no, aware, participate", "no, aware, does not participate" and "no and not aware". As described in the simulation process of question 8, "yes or no" means whether the observation profits from gaming-related platforms, "joined before or after" means whether the observation joined the platform before or after the pandemic began, "aware or not aware" means whether the observation knows about the existence of these platforms, and "participate or does not participate" means whether the observation visits or engages in such platforms.

The variables "severity_level" and "impact_of_new_normal" are numeric variables, but the numbers in these variables represent the level or extent the observation feels towards the question, and they are not actual numerical counts. The variable "severity_level" can take a value from 1 to 5, with 1 being the pandemic is not severe in the observation's area and 5 being the pandemic is very severe in the area. For the variable "impact_of_new_normal", it can also take a value from 1 to 5, with 1 being the spending habits of the observation is not impacted by the new normal at all and 5 being greatly impacted.

Numerical Summaries

In the numerical summary and plot sections of this report, the tables are created using R packages Tidyverse[17], knitr[19], and kableExtra[20], and the plots are created using R packages Tidyverse[17] and stringr[18].

For the first numerical variable, "money_diff", the mean is 28.9287209, the median is 29.3180145 and the variance is 805.5151829. Since we are concerned with the difference in money spent and the sample mean is positive, we can see that on average, there is an increase in the monthly spending on mobile in-game purchases after the pandemic happened. Note that we will be constructing a 95% confidence interval for the true mean of this variable.

When grouped by income levels, the sample shows that the mean monthly spending increases the most for observations from the highest income category, which is 39.579 CAD more per month. Simultaneously, the monthly mean for observations from the lowest income category has the least increase, which is 20.794 CAD, as seen in Table 1 below. This makes sense because it is consistent with different capacities of consumption for individuals from different income categories, where the people with higher income may afford to spend more money on mobile games.

Table 1: Difference in monthly spending on mobile game in-app purchases by income categories

income_cat	n	mean	variance
< $1000	15	20.79367	1285.8906
>= $7000	11	39.57875	667.6213
$1000 - 2400	41	32.32743	799.5700
$2400 - 3000	13	29.56854	958.4822
$3000 - 5000	77	29.73892	734.9078
$5000 - 7000	43	24.15721	774.5825

On the other hand, when grouped by the severity levels of the pandemic in the observation's area, the sample data shows that the mean monthly spending on mobile games has increased for all groups, as seen in Table 2. However, the mean spending of the two groups with the highest two severity levels has increased the most, which are 31.867 CAD and 30.783 CAD. Although people from a more severe area may be more financially affected by the pandemic, the strict quarantine happening in those areas may lead to people devoting more to mobile games. As a result, it makes sense that the higher increase in spending is seen in the observations from the areas where the pandemic is more severe.

Table 2: Difference in monthly spending on mobile game in-app purchases by severity levels

severity_level	n	mean	variance
1	41	26.38878	912.7939
2	40	27.11402	961.8759
3	43	28.70134	697.2536
4	42	31.86784	641.9171
5	34	30.78342	902.7616

The numerical variable "time_diff" has a mean of 0.2992642, a median of 0.2662396 and a variance of 8.1935593. We can see that in the sample, there is an average increase of 0.299 h in the weekly gameplay times before and after the pandemic began. Note that we will be performing a hypothesis test for the mean of this variable.

Table 3 shows the weekly gameplay time differences for groups reporting different severity levels. The sample indicates that the observations living in an area with the highest level of COVID-19 severity decrease their weekly gameplay time by 0.622 hours on average, which is the most out of the 5 groups. On the other hand, the observations reported a severity level 2 have the greatest increase in weekly gameplay time, which is 1.325 hours on average. This makes sense because the reason for an area to have median to low pandemic severity may be that they have strict local social distancing and quarantine acts that encourage people to stay inside. As a result, the people from these areas may spend more time on mobile games due to isolation and quarantine, whereas other people may be living their life as pre-COVID, and that could be why the pandemic is more severe in their areas.

Table 3: Difference in gameplay time (h) by severity levels

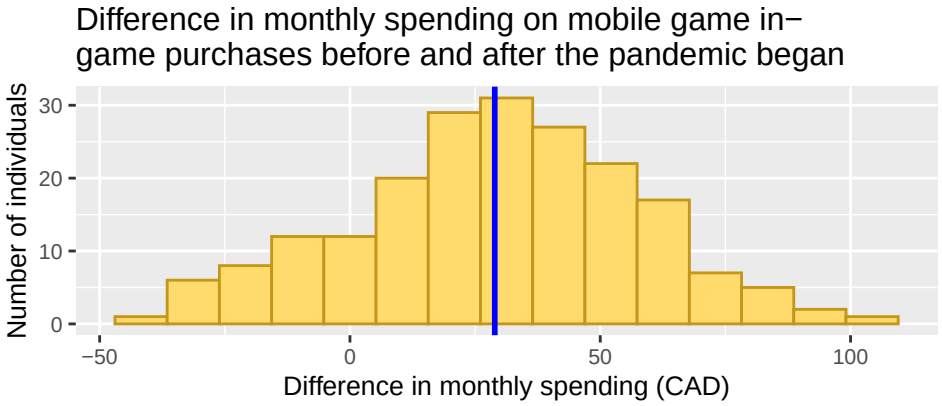
severity_level	n	mean	variance
1	41	-0.1806299	11.168159
2	40	1.3251721	9.374396
3	43	0.4636324	5.191420
4	42	0.3681059	6.912773
5	34	-0.6219078	7.120417

Interestingly, in Table 4 below, when we are only looking at male and female mobile game players in our sample, they behave differently regarding the change in the money and time spent on mobile games before and after the pandemic began. Although male mobile game players have a greater increase in the monthly spending on average in our sample, the average increase of female gamers' weekly gameplay time is greater than the male gamers, as seen in the table below. This is interesting because women are reported to spend 25% more time on mobile games than men [1], and in our sample, women's average gameplay time increase is 38% greater than men.

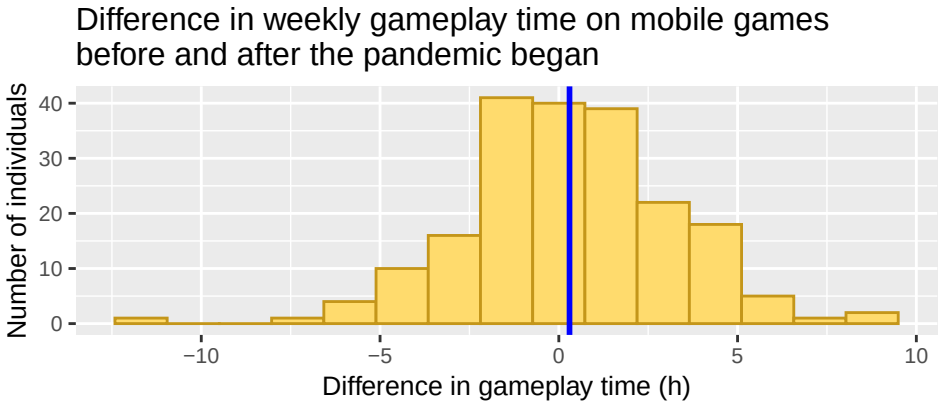
Gender	Proportion in the sample	Mean difference in money spent monthly	Mean difference in time spent weekly
Female	0.385	24.9948469	0.4227421
Male	0.54	30.9345086	0.3055661

Plots and Descriptions

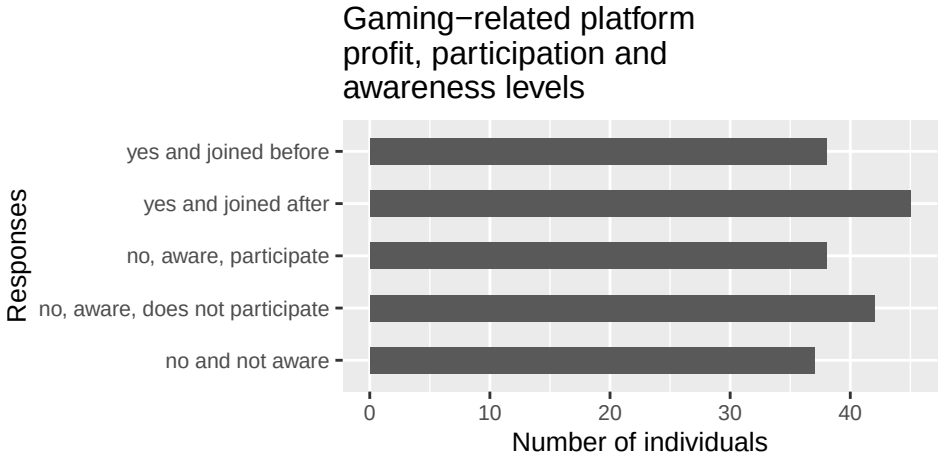
Below is a histogram of the variable "money_diff", which shows the difference in monthly spending on mobile game in-game purchases before and after the pandemic began. The vertical line in blue represents the sample mean of difference in monthly spending, which is 28.929 CAD. The data is bell-shaped and centred at the mean value, which relates to the normal distribution that the data is simulated from. A 95% confidence interval will be constructed for the mean of this variable.



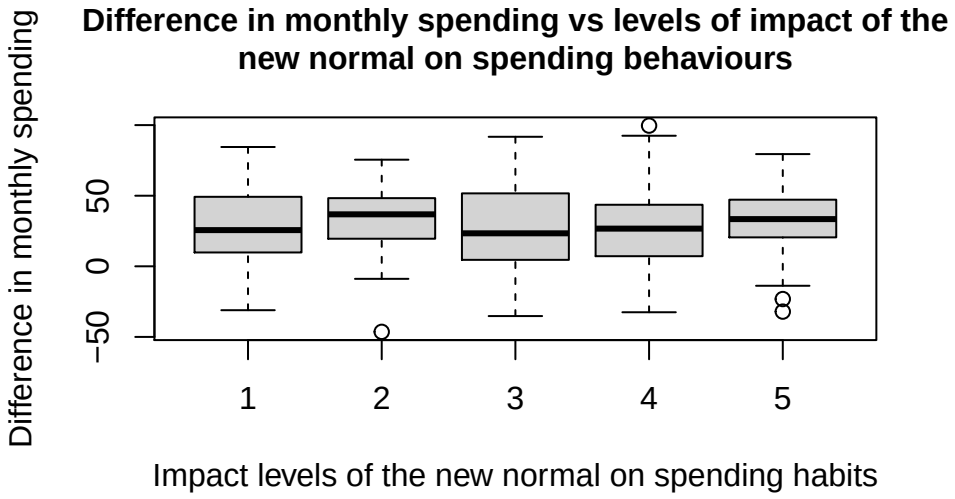
This histogram below is for the variable "time_diff", which shows the difference in weekly gameplay time spent on mobile games before and after the pandemic started. Similarly, the blue vertical line represents the sample mean, which is 0.299 hour. This histogram is slightly left skewed with some outliers lying below -10, but most observations are centred around the sample mean. A hypothesis test will be performed on the mean value of this variable.



This is a barplot for the variable "platform", indicating the observation's profit, participation, and awareness level of the gaming-related platforms. Because there is no underlying assumptions for either of the options during the simulation process, the graph appears to be generally uniform, with the category "yes and joined after" having the greatest number of individuals. This is consistent with the data reported by TwitchTracker, which indicates a 90% increase in the number of unique twitch streamers in 2020 compared to 2019 [15]. This means that a large amount of new content creators have joined these platforms after the pandemic started.



Below are boxplots of the medians for monthly spending difference grouped by the self-reported levels of impact of the new normal on the observation's spending behaviours, where 1 being not impacted at all and 5 being heavily impacted. These boxplots explore the relationship between the objective difference in spending amounts and the subjective views on the impact brought by the new normal. From the visualization, we can see that the impact level of 2 and 5 have the higher median increases in monthly spending. The highest level of impact (5) is consistent with the great increase in the spending, but people who responded the impact level of 2 appear to understate their actual average increase, which in fact has the highest median of 36.802 CAD among all groups of impact levels.



All analysis for this report was programmed using `R version 4.0.2` [10].

Methods

In this report, a hypothesis testing will be performed on the mean of variable "time_diff" and a 95% confidence interval will be constructed for the true mean of variable "money_diff". For both hypothesis testing and confidence interval, a significance level of $\alpha = 0.05$ is used.

Hypothesis Testing

The first variable of interest is "time_diff", which is the weekly gameplay time difference in mobile games before and after the pandemic began. The parameter of interest is the mean of this variable, which represents the average difference in the gameplay time. Because this variable represents a difference, or a change, the mean of this variable can be either positive, negative or zero. The goal is to investigate whether there is a change in the gameplay time or not, so we want to know if our sample mean is far from zero enough such that we are confident to say that there may be a change in the gaming time, no matter it is an increase or a decrease.

We will use a t-test [11] in the R package EnvStats [6] to perform a hypothesis testing for the mean difference of the monthly spending on mobile game in-game purchases before and after the pandemic. This is appropriate because although we simulated the data from a normal distribution with a pre-specified mean and variance, we want to treat this data as the responses gathered from an actual sampling process, so we have no assumption for the true, underlying mean and variance when constructing the confidence interval. Thus, a t-distribution with a degree of freedom of 199 is used for hypothesis test instead of a normal distribution, which is the distribution used in our simulation.

The null hypothesis is that there is no difference in the average weekly gameplay time on mobile games before and after the pandemic started.

The alternative hypothesis is that there is a difference in the average weekly gameplay time on mobile games before and after the pandemic started.

Or we can represent the hypotheses as

$$H_0 : \mu = 0$$
$$H_1 : \mu \neq 0$$

We will compute our test statistic using the Student's t-test formula [11]

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

where $\bar{x}$ represents the sample mean, μ represents the mean in the null hypothesis, s represents the sample standard deviation, and n represents the number of observations in sample.

Under the null hypothesis, the true mean is zero, because the hypothesis states that there is no difference in the gameplay time. As a result, the μ in this formula [11] is zero in this hypothesis testing. After computing the test statistic, we will refer to the t-distribution table and calculate the p-value, which represents the probability of observing data that is as extreme or more extreme than the test statistic under the null hypothesis. Because the mean could be positive or negative, we will be doing a two-sided test where probability of lying in both tails is included in the p-value. If the calculated p-value is greater than our significance level of 0.05, then we fail to reject our null hypothesis that there is no difference in the weekly gameplay time in mobile games before and after the pandemic began. However, if p-value is smaller than the significance level, then we are able to reject the null hypothesis.

Note that a 95% confidence level for the true mean of the difference in weekly gameplay time will also be derived by the method we use [6] as additional information.

Confidence Interval

The second variable of interest is "money_diff", which represents the difference in the spending on in-game purchases in mobile games per month before and after the pandemic began. The parameter of interest is the mean of this variable, which is the mean difference in spending. We aim to see the average change in people's spending on mobile games, so the difference can be either positive, negative, or zero. Thus, we want an interval that we can be confident in and that it captures the true mean of the difference in monthly spending for most of the time.

To achieve this, we will derive the 95% confidence interval (CI) for the true mean of the difference in the monthly spending (in CAD) on mobile in-game purchases before and after the pandemic began, using a t-distribution with a degree of freedom of 199. Performing CI using a t-distribution is appropriate because, like the variable "money_diff", we want to treat the true, underlying mean and variation as unknowns even though we simulated this data from a normal distribution where we know the mean and variance. As a result, we assumed we do not know the true, underlying mean and variance when constructing the confidence interval, and thus we use a t-distribution again instead of a normal distribution.

The formula used is [13]

$$\bar{x} \pm t_{\frac{\alpha}{2}, df=n-1} \frac{s}{\sqrt{n}}$$

where $\bar{x}$ represents the sample mean, s represents the sample standard deviation, and n represents the number of observations in sample.

After plugging the values from our sample into the formula [13], we will obtain the upper and lower bounds for our interval. And then we will be able to say that we are 95% confident that this interval captures the true mean of the difference in monthly spending on mobile games.

To confirm with the calculated 95% confidence interval, we will use a t-test [6][11] using the same procedure as the hypothesis testing above. This test will produce the 95% confidence interval, as well as additional information such as test statistic and p-value. The test statistic and p-value are calculated under the null hypothesis that there is no difference in the average of the monthly spending on mobile game in-app purchases before and after the pandemic.

Results

The Table 5 below is a summary of the results for the hypothesis testing and confidence interval using t-test [6][11].

Variable	Sample mean	Test statistics	Degree of freedom	p-value	95% Confidence interval
time_diff	0.2992642	1.4785	199	0.1408	(-0.09986974, 0.69839811)
money_diff	28.92872	14.415	199	$<$ 2.2e-16	(24.97124, 32.88621)

For the hypothesis testing of mean time difference, the resulted p-value is 0.1408, which is greater than our significance level 0.05. Alternatively, we can see that the mean given by the null hypothesis, which is 0, falls within the 95% confidence interval of (-0.09986974, 0.69839811). As a result, we fail to reject the null hypothesis that there is a difference in the weekly gameplay time in mobile games before and after the pandemic began. This means that the time difference in our sample is not substantial enough to be considered as significant.

The result is reasonable and consistent with the way we simulated the data for weekly gameplay time before and after the pandemic. The sample mean of the time difference is 0.2992642 hour, which is relatively small yet still positive. The average time spent on video games in U.S. in 2019 and 2021 is used to simulate our data, but the increase in the average time is relatively small (0.1 h), as reported by Limelight Networks [7] [8]. However, this result does not reflect the proportion of gamers increasing their mobile gameplay time,

which is 63% of the participants as reported by Ward [16]. As a result, although we fail to say that there is actually a statistically significant difference in the duration of gameplay time based on our sample, we still observed a proportion of 0.545 of the observations that have increased their mobile gameplay time in our sample.

On the other hand, the 95% confidence interval for the true mean of the difference in the monthly spending on mobile in-game purchases before and after the pandemic is (24.97124, 32.88621). This 95% confidence interval is confirmed by both manual calculation and the t-test using the R package EnvStats [6]. This means that we are 95% confident that the true, population mean of monthly spending difference falls in this interval. Or in other words, we are 95% confident that in the population, the average difference of monthly spending is somewhere between 25 CAD and 33 CAD, which indicates a substantial average increase in the money spent on mobile games. Moreover, if we were to do a hypothesis testing with the null hypothesis that there is no difference in the monthly spending before and after the pandemic, then the p-value that is smaller than 2.2e-16, which is also much smaller than our significance level of 0.05. This indicates that we can reject our null hypothesis that there is no difference in the monthly spending on mobile games.

This result seems reasonable because according to Coulson (2021), there is an increase of 23 USD (29.29 CAD) in the monthly spending on mobile games for iPhone users in 2020 compared to 2019 [4]. The value 29.29 lies within the confidence interval (24.97124, 32.88621) as well. Simultaneously, as reported by the NPD group, the spending on mobile games in US has increased by 5% since the second quarter of 2020 [14], indicating the mobile game consumers continue to spend more in 2021 with lots of new consumers joining the market during the pandemic.

However, this result may be biased due to the limitations in the simulation process. Because most of the data used for the mean and variance of relevant distributions are from different sources, there will be underlying time, area, and demographic differences in the referenced literature, and these different reports all have limitations and biases in themselves. As a result, there may be inherent bias within the simulation process itself, leading to potentially biased results. Also, as the information obtained from the literature are mostly from the U.S., it may be more appropriate to generalize the results of this report to the mobile gamers in the U.S. than Canada. Therefore, we need to be careful about what results we are able to generalize and how we can generalize these results to the mobile game players in Canada.

In conclusion, we are able to construct a 95% confidence interval for the true mean of the difference in monthly spending on mobile games, which is (24.97124, 32.88621), but we fail to reject the null hypothesis that there is no difference in the average weekly gameplay time on mobile games before and after the pandemic. Hopefully, this report can provide some interesting analysis and insights into how COVID-19 has changed the way people play mobile games, and how that may be related to various other environmental, economic, social and psychological factors brought by the global pandemic. By exploring the curious dynamics between the pandemic and people's responses, we eventually may better adapt to this new normal moving forward.

Bibliography

1. Andre, L. (n.d.). *55 mobile Gaming Demographics Statistics: 2021 data on market share & spending.* FinancesOnline. https://financesonline.com/mobile-gaming-demographics/. (Last Accessed: September 28, 2021)

2. Clement, J. (2021, July). *Global game developer gender 2021.* Statista. https://www.statista.com/statistics/453634/game-developer-gender-distribution-worldwide/. (Last Accessed: September 28, 2021)

3. Clement, J. (2021, July 14). *UK mobile gaming by age 2021.* Statista. https://www.statista.com/statistics/300522/mobile-gaming-in-the-uk-by-age/. (Last Accessed: September 28, 2021)

4. Coulson, J. (2021, April 13). *The average iphone user spent more than $75 on games in 2020.* TheGamer. https://www.thegamer.com/average-iphone-spend-games-2020/.(Last Accessed: September 28, 2021)

5. Kominers, S.D. (2021, April 17). *How the pandemic revealed the power of live video.* Bloomberg. https://www.bloomberg.com/opinion/articles/2021-04-17/livestreaming-surged-during-the-pandemic. (Last Accessed: September 26, 2021)

6. Millard, S.P.(2013). *EnvStats: An R package for environmental statistiscs.* Springer.

7. Limelight Networks. (2019). *Market research: The state of online gaming – 2019.* Limelight Networks. https://www.limelight.com/resources/white-paper/state-of-online-gaming-2019/. (Last Accessed: September 28, 2021)

8. Limelight Networks. (2021). *State of online gaming 2021.* Limelight Networks. https://www.limelight.com/lp/state-of-online-gaming-2021/. (Last Accessed: September 28, 2021)

9. Statistics Canada. (2021, March 23). *Table 11-10-0239-01 Income of individuals by age group, sex and income source, Canada, provinces and selected census metropolitan areas. Statistics Canada.* https://doi.org/10.25318/1110023901-eng. (Last Accessed: September 28, 2021)

10. R Core Team (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/. (Last Accessed: September 29, 2021)

11. Student. (1908). The probable error of a mean. *Biometrika*, 1–25.

12. *Tables.* (n.d.). R Markdown. https://rmarkdown.rstudio.com/lesson-7.html. (Last Accessed: September 29, 2021)

13. *T confidence interval for a mean.* (n.d.). R. https://cran.r-project.org/web/packages/distributions3/vignettes/one-sample-t-confidence-interval.html. (Last Accessed: September 29, 2021)

14. The NPD Group. (2021, July 22). *The NPD Group: Second Quarter 2021 U.S. consumer spending on video game products Increased 2% to $14 billion.* NPD. https://www.npd.com/news/press-releases/2021/the-npd-group-second-quarter-2021-u-s-consumer-spending-on-video-game-products-increased-2-to-14-billion/. (Last Accessed: September 26, 2021)

15. *Twitch statistics & charts.* (n.d.). TwitchTracker. https://twitchtracker.com/statistics. (Last Accessed: September 28, 2021)

16. Ward, L. (2021, May). *What mobile gaming's "new normal" should look like after the COVID-19 pandemic.* IDC. https://www.idc.com/getdoc.jsp?containerId=US47730721. (Last Accessed: September 26, 2021)

17. Wickham et al., (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686. (Last Accessed: September 29, 2021)

18. Wickham, H. (2019). *stringr: Simple, consistent wrappers for common string operations.* R package version 1.4.0. https://CRAN.R-project.org/package=stringr. (Last Accessed: September 29, 2021)

19. Xie, Y. (2021). *knitr: A general-purpose package for dynamic report generation in R.* R package version 1.31.

20. Zhu, H. (2021). *kableExtra: construct complex table with 'kable' and pipe syntax.* R package version 1.3.4. https://CRAN.R-project.org/package=kableExtra. (Last Accessed: September 29, 2021)

Appendix

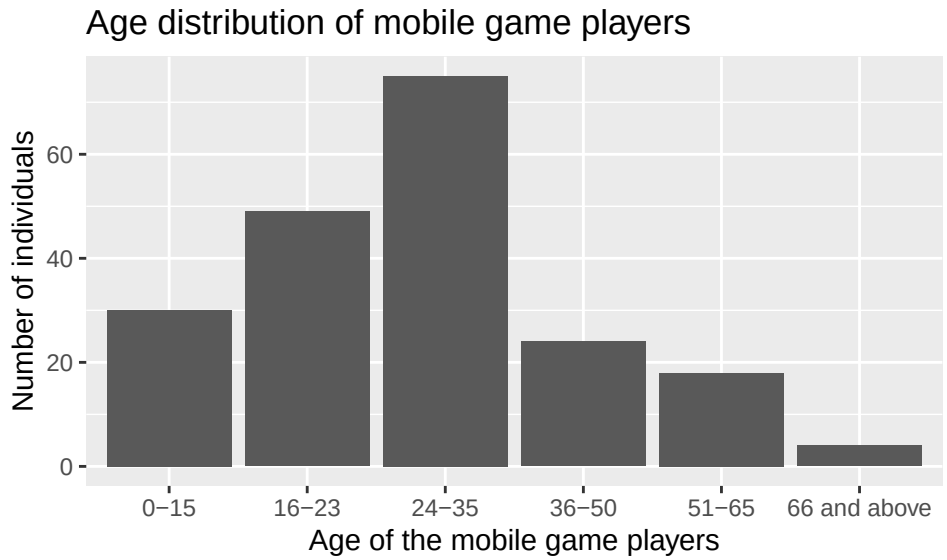
Here is a glimpse of the data set simulated with 11 variables corresponding to the 11 quesitons on the survey:

```
## Rows: 200
## Columns: 11
## $ age                  <ord> 24-35, 16-23, 66 and above, 36-50, 24-35, 16-23, ~
## $ gender               <fct> female, male, male, male, male, female, female, s~
## $ income               <dbl> 3202.2962, 2856.2306, 1098.2610, 4988.1470, 4551.~
## $ money_before         <dbl> 41.66809, 85.05435, 70.86991, 72.62691, 55.88510,~
## $ money_after          <dbl> 83.03561, 113.03151, 109.09639, 99.99045, 102.253~
## $ time_before          <dbl> 3.263757, 1.647634, 4.212104, 7.968045, 8.319578,~
## $ time_after           <dbl> 4.202864, 10.598535, 4.014041, 5.943019, 5.557787~
## $ platform             <fct> "no and not aware", "yes and joined after", "yes ~
## $ severity_level       <int> 4, 2, 3, 4, 5, 3, 4, 5, 5, 1, 2, 1, 2, 5, 2, 2, 5~
## $ impact_of_new_normal <int> 5, 4, 5, 1, 2, 1, 2, 3, 4, 2, 1, 3, 2, 5, 4, 1, 2~
## $ cont_spending        <int> 1, 2, 2, 3, 5, 3, 4, 3, 1, 1, 2, 5, 5, 2, 1, 2, 3~
```

Here is a glimpse of the cleaned data set with 9 relevant variables to this report:

```
## Rows: 200
## Columns: 9
## $ age                  <ord> 24-35, 16-23, 66 and above, 36-50, 24-35, 16-23, ~
## $ gender               <fct> female, male, male, male, male, female, female, s~
## $ income_cat           <chr> "$3000 - 5000", "$2400 - 3000", "$1000 - 2400", "~
## $ money_diff           <dbl> 41.367523, 27.977153, 38.226483, 27.363536, 46.36~
## $ time_diff            <dbl> 0.93910657, 8.95090146, -0.19806236, -2.02502622,~
## $ platform             <fct> "no and not aware", "yes and joined after", "yes ~
## $ severity_level       <int> 4, 2, 3, 4, 5, 3, 4, 5, 5, 1, 2, 1, 2, 5, 2, 2, 5~
## $ impact_of_new_normal <int> 5, 4, 5, 1, 2, 1, 2, 3, 4, 2, 1, 3, 2, 5, 4, 1, 2~
## $ cont_spending        <int> 1, 2, 2, 3, 5, 3, 4, 3, 1, 1, 2, 5, 5, 2, 1, 2, 3~
```

Below is a barplot [17] for the age distribution of participants in the sample, which is right-skewed and the category with the greatest number of individuals is 24-35. This is reasonable because "25-34" is reported to be the largest age group among all mobile gamers in the U.S.[1].



The table [17][19][20] below is the difference in monthly spending on mobile games before and after the pandemic started, grouped by how likely the individual will continue spending on the in-game purchases in the future. This is interesting because in the sample, individuals who responded the least likely to spend in the future (level 1) actually have the greatest average increase in the monthly spending after the pandemic started, whereas people who have less increase in the spending are more likely to keep spending. This may be explained by the irrational spending on games that could happen during the pandemic lockdown, where individuals may be regretful and decide not to spend on mobile games in the post-pandemic future.

Table 6: Difference in monthly spending by likelihood of continuing spending

cont_spending	n	mean	variance
1	51	33.06287	854.7231
2	32	31.35975	557.7734
3	39	30.88958	917.0471
4	34	24.09760	756.0102
5	44	24.36396	873.3293