STA304 - Fall 2022

Assignment 1

Zhaoqi Li - 1006324639

September 29th 2022

Part 1

Goal

With the popularity of smart phones and social media, online shopping has become a new trend, especially for university students. Online shopping provides lot of convenience to people, which saves transportation costs and shopping time to a great extent. According to the article "The latest online shopping trends to hit university campuses", Purvis (2020) expresses that online shopping is in a booming stage and the university students is the largest group of using online shopping. (Purvis, 2020) Therefore, university student is an important research group for online shopping. My topic will mainly analysis what factors affect university students shop online. I will analysis university students' online shopping situation under different factors (like gender, stratification, etc.) to analyze which aspects affect university students' online shopping. This survey researches the online shopping information of UofT students. It includes the consumption, frequency, gender, grades, etc. which helps us to analyze what will influence students' online shopping from a variety of factors. For example, I can calculate the mean consumption of male and female to compare the difference.

Procedure

This project is to investigate the online shopping of all students in the University of Toronto. Therefore, my target population is all students who current study in University of Toronto. Through three years' study at the University of Toronto, I have met many classmates in different courses. I have a list of UofT students' social media accounts from WeChat and Instagram. Therefore, the frame population is the current students at University of Toronto who leave the social media accounts with me. I have about 500 social media accounts from WeChat and Instagram. I will post my survey link through these social media and my UofT classmates volunteered to fill out the survey. Therefore, the sampling population is the UofT students who respond the survey. It does not involve random selection, so it is non-probability sampling. As students volunteered to participate in the survey, the sampling method is volunteer-base sampling.

Volunteer-base sampling is a convenience way to get responses, which can collect the date simply and quickly. As people volunteer to participate the survey, we do not need to pay them. Therefore, this method also can reduce the cost for research. However, as a non-probability sampling method, volunteer-base sampling is self-selected by participants rather than random selection. The participants may have strong views on the topic, and their responses would tend to be extreme. This will lead to biased and unreliable results (Voxco, 2021). Although volunteer-base sampling is easy to collect data and saves time and money, it is not applicable in most research, because the data collected by this method is not very accurate and representative.

Showcasing the survey.

This is my survey question link: https://forms.gle/NvXGdhtHfE7dvem4A

Question 1 How many times do you shop online monthly? Please enter a number.

This question is a numerical question which mainly investigates the frequency of university students' online shopping. Therefore, this question can be used to analyze the relationship between university students' online shopping frequency and online shopping expenses. This is a short and simple question, and the participants are easy to understand. For this question, I need numerical variable from the participants, so I remind them enter a number to make sure I can get the applicable data. However, this question has some limitations, which requires participants to recall their online shopping times in a month. It may lead to some errors because some people may not remember their online shopping times.

Question 2 How much do you spend on your online shopping monthly (the unit is CAD)? • Under $100

- $100 - $499
- $500 - $999
- $1000 - $1500
- Over $1500

This question is a categorical question, which divides the amount of university students' online shopping consumption into five intervals. Therefore, we can know the approximate interval of college students' online shopping consumption level, and then analyze what factors mainly affect the consumption level. This question is a multiple-choice question, and I haven't set too many options, so it is easy for participants to choose. Additionally, the choices of multiple-choice questions should be mutually exclusive, so that the participants can make correct responses without ambiguity (Smith and Fisher, 2018). Therefore, to avoid ambiguity, each option is differentiated, and there are no mutually exclusive options. I also mention the monetary unit which made the participants understand the question clearly. The defect of this question is that the numerical interval is relatively large, which will cause some errors.

Question 3 How satisfied are you with online shopping?

Level 1-5 (dissatisfied to satisfied)

This question is a categorical question, which includes five satisfaction levels of online shopping, with 1 being the least satisfied and 5 being the most satisfied. This is a Likert scale question, which is simple and reliable in reduce survey bias (Smith and Fisher, 2018). Through this question, we can know participants' satisfaction with online shopping and get the relationship between university students' satisfaction with online shopping and spending by analyzing the comparison with spending. This is a simple and short question. I clearly label the interval of satisfaction, so the participants are easy to understand the question. Through the results of the survey, I found that participants usually avoid extreme options, such as the most satisfied and the least satisfied. This will make the results concentrated in the middle area, which may not fully represent the opinions of the participants themselves, thus causing some deviation in the results.

Part 2

Data

My data comes from the survey of students currently studying at UofT. I want to investigate the online shopping situation of UofT students, the information obtained from the survey is more representative.

I collect my data through social media. I have about 500 social media accounts of UofT students, which come from WeChat and Instagram respectively. I asked students to fill in the survey voluntarily by posting my survey link. I received 47 responses in two days. The survey is filled out voluntarily, some people may do not want to reply and there will be non-response bias. Additionally, those who volunteer to fill out the questionnaire are those who are interested in this topic, so the results may be biased. I downloaded the survey into csv format. Each row represents a person with her/him answers, and each column represents the questions I asked. Finally, I get 2 numerical variables and 6 categorical variables.

Table 1: Online Shopping Situation of UofT Students

gender	grade	course_number	frequency	spending	package_lost	satisfaction
Female	Third Year	5	8	$500 - $999	No	4
Male	Second Year	4	4	$100 - $499	Yes	3
Male	First Year	5	3	$100 - $499	No	4
Female	Third Year	6	10	$1000 - $1500	No	4
Female	Second Year	4	3	$100 - $499	Yes	2
Male	Fourth Year	3	1	Under $100	No	3

I import the data and store it in the dataset called “my data”. According to the survey, I have 9 questions which are all sentences. Therefore, I rename the sentences as a word which will be easy and clear. Additionally, there is an open-ended question “reason” in my survey. I remove the variable “reason” in my dataset. According to the survey question, there is a question (the category of products) which can choose multiple options. This is a categorical question and not easy to analyze, so I also remove the category of the product. Then, I remove all missing data which are the no-responses questions.

The important variables I will analysis are below:

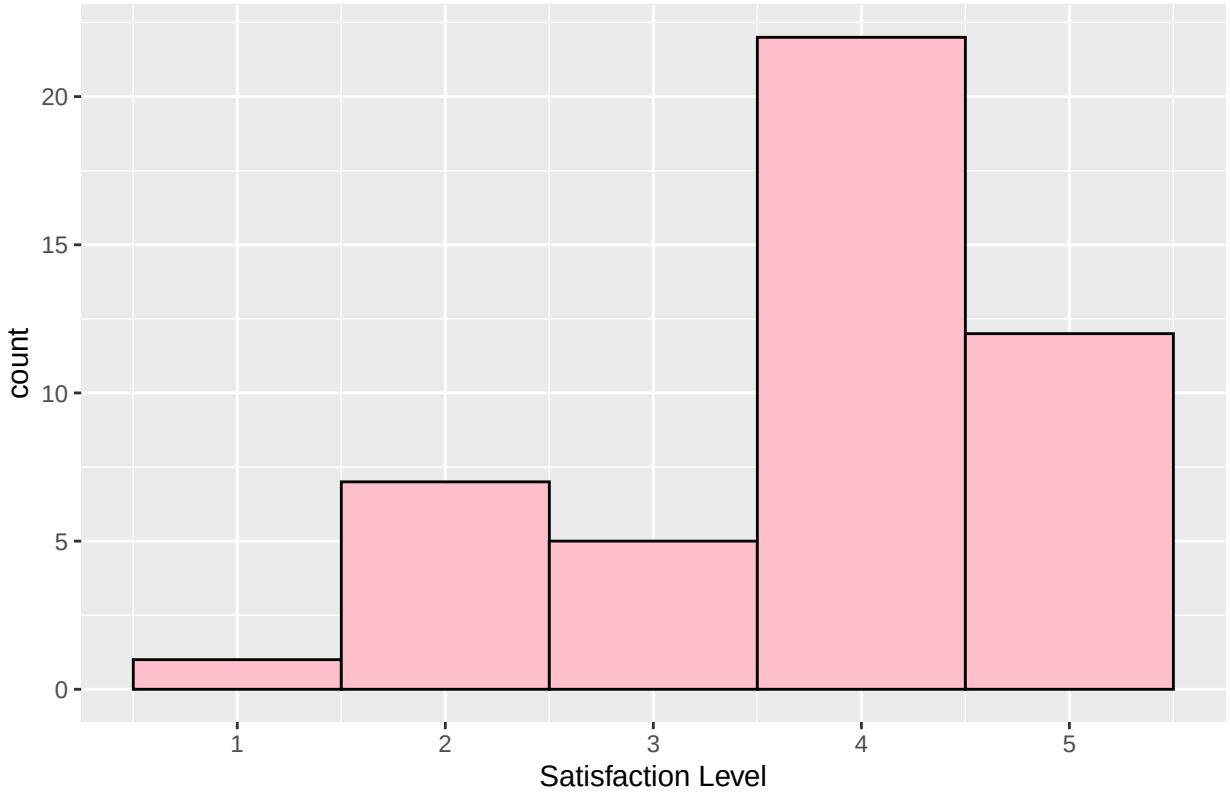
- Frequency: It expresses the sum of times for people shop online monthly, which is a numerical variable.
- Spending: It means the spending of online shopping monthly, which expresses the 5 spending intervals. This is a categorical variable.
- Package lost: This is the situation of lost package or not. It expresses if you have ever lost a package which has two options: yes and no. This is a categorical variable.
- Satisfaction: it expresses the satisfaction level of online shopping, which includes 5 levels from 1 to 5(dissatisfied to satisfied). This is a categorical variable.
- Gender:Sex of biological classification, male or female. This is a categorical variable.

Table 2: Numerical Summaries For Data Frame

frequency	satisfaction
Min. : 1.000	Min. :1.000
1st Qu.: 4.000	1st Qu.:3.000
Median : 7.000	Median :4.000
Mean : 5.957	Mean :3.787
3rd Qu.: 8.000	3rd Qu.:4.500
Max. :12.000	Max. :5.000

I summarized two numerical variables which are frequency and satisfaction. For frequency, it expresses the sum of times for UofT students shop online. The mean is 5.957 which means on average every students shop online 5.957 times monthly. The frequency of online shopping ranges from 1-12 times, with min 1 and max 12. For satisfaction, it expresses the satisfaction level for students shop online which is from 1 to 5. According to the table 1, the mean is 3.787, which means the average satisfaction of online shopping for UofT students is 3.787. The satisfaction of online shopping ranges is from level 1 to level 5(min is 1 and max is 5).

Figure 1:The Satisfaction Level of UofT Students Shop Online



This histogram expresses the satisfaction level of UofT students shop online which is from 1 to 5. According to the histogram, the center is 4 and most people are concentrated in the bin 4 to 5 which means most people are satisfied for online shopping. The shape of the histogram is left-skewed.

This side-by-side boxplots compared the frequency of female and male shop online monthly. According to the boxplots, the median of females' monthly online shopping frequency(about 7 times) is higher than the median of males' monthly online shopping frequency(about 5 times). Therefore, females may shop online more frequently than males.

Methods

HT is Hypothesis Test which means what extent of the data can support a statistical statement by using the independent distribution. There are two hypotheses: null hypothesis(H0) and alternative hypothesis(H1). According to Hypothesis Test, we can get a probability value to express whether it support our statistical statement or not. The p-value expresses if the null hypothesis is true, it observes the more extreme probability (Bevans,2020). In this project, I use two-sample hypothesis test.

According to figure 2, I use side-by-side boxplots expresses the frequency of online shopping by male and female. Therefore, I want to research whether the average online shopping frequency between men and women is equal. I assume these two variables continuous and independent. The sample variances are similar, and the population follow the normal distribution.

Null Hypothesis & Alternative Hypothesis

Null Hypothesis: The average online shopping frequency of male is equal to the average online shopping frequency of female. Alternative Hypothesis: The average online shopping frequency of male is different to the average online shopping frequency of female.

$$H_0 : \mu_M = \mu_F \quad \text{vs.} \quad H_1 : \mu_M \neq \mu_F$$

Test statistic

$$T = \frac{(\bar{X} - \bar{Y}) - 0}{S^2\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

95% Confidence Interval

The 95% confidence interval means I am confident that 95 out of 100 times the estimated value will fall between the give range of value(Bevans, 2020). I will use 95% confidence interval to estimate the interval between online shopping frequency of males and females.

$$95\%CI : [\bar{x} - t_{n-1}\frac{s}{\sqrt{n}}, \bar{x} + t_{n-1}\frac{s}{\sqrt{n}}]$$

Results

Table 3: Summary of Statistic Test

Variance.of.Male	Variance.of.Female	P.value
8.277056	5.193333	0.01201

Table 4: The Conference Interval of Male

CI.Upper.Bound	CI.Lower.Bound	Significance.Level	gender
10.54788	-0.7297007	0.05	Male

Table 5: The Conference Interval of Female

CI.Upper.Bound	CI.Lower.Bound	Significance.Level	gender
11.34654	2.41346	0.05	Female

According to the table 3, the variance of male and variance of female is similar. The p-value of hypothesis test is 0.01201 which is large than 0.01 and less than 0.05. Therefore, we have strong evidence against null hypothesis which means the average online shopping frequency of male is not equal to the average online shopping frequency of female.According to table 4 and 5, I get the 95% confidence interval of males' online shopping frequency is between -0.73 to 10.55 times, and the 95% confidence interval of females' online shopping frequency is between 2.41 to 11.35 times. Therefore, the frequency of online shopping for females per month is higher than that for males per month. We can conclude that females like online shopping more than males in University of Toronto.For these two result, the confidence interval is wide which means the estimate is not very precise, because my sample size is not large enough. In order to get more accurate results, my sample size should be larger.

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. Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how.* Springer Science & Business Media.
3. Allaire, J.J., et. el. *References: Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: May 5, 2021)
4. Purvis, A. (2020, November 3). *The latest online shopping trends to hit university campuses* Quadient [https://www.quadient.com/blog/latest-online-shopping-trends-hit-university-campuse]
5. Voluntary Response Sample. (2021, April 20). Retrieved from Voxco website: https://www.voxco.com/blog/voluntary-response-sample/
6. Online shopping survey of UofT students. (n.d.). Retrieved October 6, 2022, from Google Docs website: https://forms.gle/NvXGdhtHfE7dvem4A
7. Tables. (n.d.). Retrieved from rmarkdown.rstudio.com website: https://rmarkdown.rstudio.com/lesson-7.html
8. Bevans, R. (2020, July 16). Understanding P-values | Definition and Examples. Retrieved from Scribbr website: https://www.scribbr.com/statistics/p-value/
9. Bevans, R. (2020b, August 7). Confidence Interval | Definition, Formulas, Examples. Retrieved from Scribbr website: https://www.scribbr.com/statistics/confidence-interval/
10. Bevans, R. (2020b, July 17). Test statistics | Definition, Interpretation, and Examples. Retrieved from Scribbr website: https://www.scribbr.com/statistics/test-statistic/
11. Tidyverse. (2019). Retrieved from Tidyverse.org website: https://www.tidyverse.org/packages/

Appendix

Here is a glimpse of the data set surveyed:

```
## Rows: 47
## Columns: 9
## $ gender           <chr> "Female", "Male", "Male", "Female", "Female", "Male",~
## $ grade            <chr> "Third Year", "Second Year", "First Year", "Third Yea~
## $ course_number    <dbl> 5, 4, 5, 6, 4, 3, 5, 5, 5, 4, 6, 5, 5, 6, 6, 4, 5, 6,~
## $ frequency        <dbl> 8, 4, 3, 10, 3, 1, 6, 2, 4, 2, 10, 7, 4, 6, 8, 3, 4, ~
## $ product_category <chr> "Household goods (laundry detergent, toilet paper, et~
## $ spending         <chr> "$500 - $999", "$100 - $499", "$100 - $499", "$1000 -~
## $ reason           <chr> "very convenient", "save money", "Save time", "I do n~
## $ package_lost     <chr> "No", "Yes", "No", "No", "Yes", "No", "No", "Yes", "N~
## $ satisfaction     <dbl> 4, 3, 4, 4, 2, 3, 4, 2, 3, 3, 4, 5, 4, 5, 4, 2, 4, 5,~
```