STA304 - Fall 2022

Assignment 1

Yihang Luo - 1006871100

2022-09-27

Part 1

Goal

The cost of college is very expensive, especially for college students, a group that does not yet have the ability to earn money. That's why the concept of saving money and money-saving habits are very important. Besides rent, the most important expense of college students may be on food. Although most college students would agree that eating well is an important part of a healthy lifestyle, they often struggle to do so. Fast food, the low intake of vegetables and fruits, and erratic dietary patterns are more indicative of college students' eating habits, and only part of them would choose the healthy food[1]. By studying in how much a college student spend on food per month, as well as their food preference, ways and meal frequency and all sorts of food related questions, we can understand whether planning healthy meals in advance can help students save money in college. Through the deeper understanding of the additional benefit on saving money for college students from having healthy dietary habits, we can advise them to pursue healthy dietary plans.

Procedure

Since I am a junior college students, it is very convenient for me to work out the survey about college students. And it is also interesting to know about the food consumption concept of people who are as me. The population target is to all college students at the same age as me as a junior. The sample is collected using the volunteer sampling from classmates in sta304 and other courses in uoft I knew, which are mostly international students around me in Canada. The frame is the list of my classmates who do the survey and follow up with all their answers to the questions.

While I save a lot of energy and time from volunteer sampling, there is a volunteer bias in the survey because I cannot attach to other college students that are not my friends or do not want to finish the survey. And since most of my friends are international students as me, the result may tend to be more effective in the international college students group instead of the total college students group.

Showcasing the survey.

https://forms.office.com/r/gxt2Pvv2WN [1]

Question 2 What is your monthly living expenses around? (without rent)

This is a numerical question to know about the total living expenses of a college student in a month. We need this data mainly because only combine this to the monthly food costs proportion to living expenses, we

can better work out the food consumption we like instead of directly asking participants for their monthly food cost. Seldom people can remember the exact cost on food. The amount of money from 750 to 2250 is appropriate for a college student's normal expense as most data set around it. There are numbers fluctuate from 750 to 2250, so the participants have a large scale to do the option. While it is too restrict maybe for the number to be fix value instead of scope so the result may show less accuracy with data, and there may be value much less than 750 or much more than 2250.

Question 3 What is your monthly food costs as a proportion of living expenses?

This is a numerical question to know about the proportion to the total living expenses we just investigate. Regard each approximately 20% part of 100% possible, each option <20%, 30%, 50%, 70%, >80% is being choose. The use of "<", ">", and multiple level of proportion let people easy to choose. While participants may be confused since the level for choosing is quite fix to a certain value instead of a scope. It is also harder to calculate data with "scope"<20%" or ">80" instead of accurate number.

Question 4 Your dietary preference is

This is a categorical question to know about the main consumption type on food, such as healthy food, fast food or non-preference food. Even though we only need to know whether students' dietary are healthy or not, multiple options could let people make their decisions closer to their real mind. So the 3 basic options for college students are appropriate and there is also an option for "else" so all kinds of participants can make a choice. Through this question, we can associate the food consumption with dietary patterns and preference and know about healthy food style or not each participants have, so we can figure out an approximate picture about the topic. While the third option "whatever is full" may be vague to some participants, and there is no option for both two options. The option would be more reasonable if it becomes a multiple-choice option.

Part 2

Data

The STA304 Questionnaire data comes from 36 people who are classmates as the same grade as me. There are 25 male and 24 female and one "prefer not to say" gender participants in the data. It is about how much a college student spend on food per month difference with various factors such as dietary preference, meal frequency and other related factors.

Common concerns about healthy eating are time constraints, unhealthy snacking, convenient high-calorie foods, stress, high prices of healthy foods, and easy access to junk food. However, factors that promote healthy behaviors are actually increased food knowledge and education, meal planning, participation in food preparation, physical activity, and the consumption of less money due to eating regularly and consuming less money. To persuade more college students to have healthy diet plans, what I hope the study will show is that a bad diet doesn't save money, but even consumes it.

I first clean the data for the survey questions. Since each column represents a question which is long and useless, I rename each column into a variable name for easier use. Then I begin to keep data that is important for me. For variable dietary preference, I only care about the food that is usually had by participants is healthy or not, so I categorize them again into two group, only "healthy" and "unhealthy" for the other types that are not healthy, so we can directly see the the data we need, and better figure out the relationship between dietary preference and food costs of college students. After observing the data, I found that only 2 participants have their monthly food costs proportion smaller than 20% or greater than 80%, only 1 participants have meals more than 3 a dayso I delete those outliers. I also change the monthly food costs proportion from percentage to decimal numbers, so we can calculate the actual food consumption more easily. Finally, I create a new variable food consumption which is from the food costs proportion on living expenses and monthly living expenses, a numerical variable where our future calculation and tests based on.

Table 1: Numerical Summary of Food Consumption

dietary_preference	n	min	max	mean	standard
Healthy	19	375	1125	632.8947	192.3994
Unhealthy	28	225	1225	658.9286	278.5723

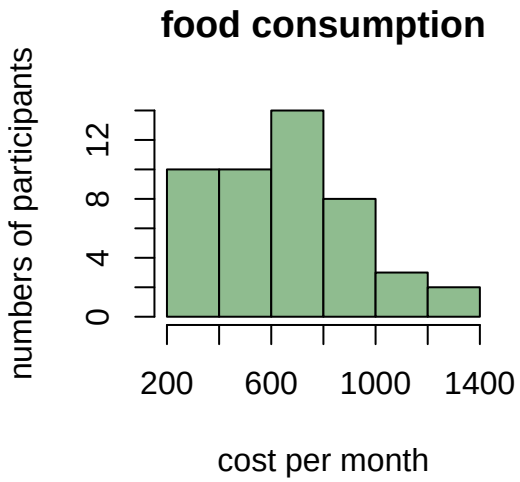
Table 2: Numerical Summary of Food Consumption

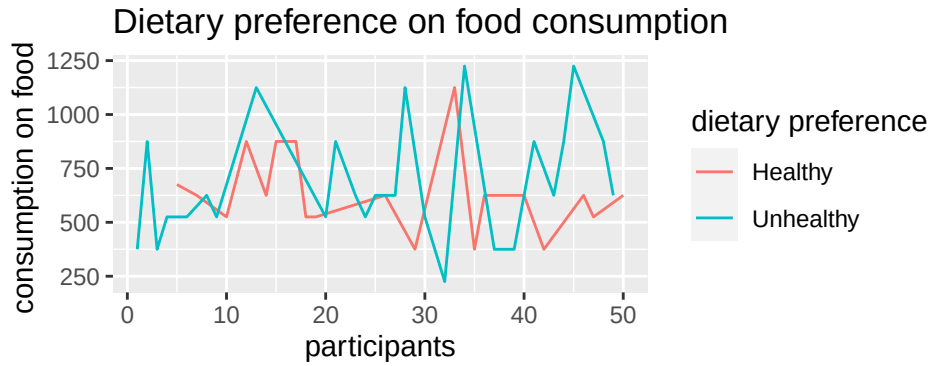
meal_frequency	n	min	max	mean	standard
2 meals	31	225	1225	626.6129	249.4941
3 meals	16	375	1125	690.6250	239.2480

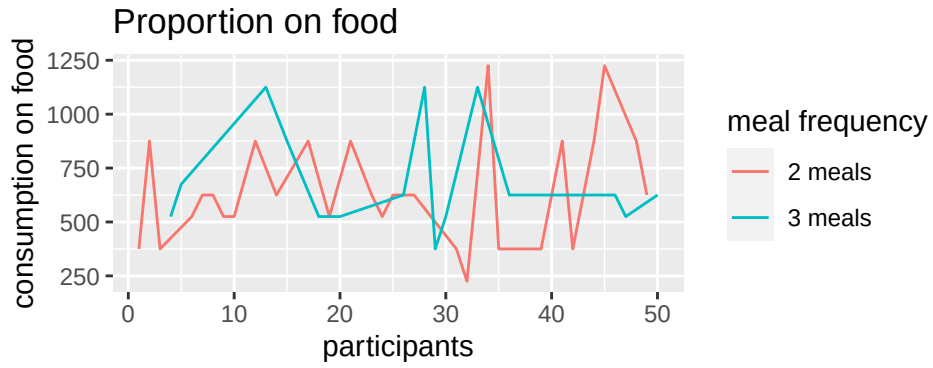
[3]From table 1, we found that there are 19 people choose healthy diet and 28 people choose unhealthy diet. So it seems that more college students choose unhealthy food for their eating habits. The spread of costs on healthy diet is from 225 to 1225, while the spread of costs on unhealthy diet is from 375 to 1125. The mean of college student's monthly costs on 2 meals is 632.8947, while the mean of college student's monthly costs on unhealthy diet is 658.9286, which is higher than that on healthy diet. The standard deviation of costs on healthy diet is 192.3994, and the standard deviation of costs on healthy diet is 278.5723, so the costs on unhealthy diet among college students varies in higher volatility.

From table 2, we found that there are 31 people choose to have 2 meals a day and 16 people choose to have 3 meals a day. So it seems that college students are more likely to have 2 meals than 3. The spread of costs on 2 meals is from 225 to 1125, while the spread of costs on 3 meals is from 225 to 1225. The mean of college student's monthly costs on 2 meals is 626.6129, while the mean of college student's monthly costs on

3 meals diet is 690.6250, which is higher than that on 2 meals. It is reasonable because consume on one more meal would certainly cost more. The standard deviation of costs on 2 meals is 249.4941, and the standard deviation of costs on 3 meals is 239.2480, so both of the costs on 2 or 3 meals among college students varies in a similar scale.







The first histogram demonstrates the food consumption of sample college students per month. We can see 600~800 dollars on food consumption is the most cost of college students. There is a skewed to the right trend to the graph, which means smaller group of people tend to consume more than 600~800 dollars instead of cost 600~800 fewer on food.

The second categorical line plot reflects participants' monthly consumption proportion on food with two groups "healthy" in red line or "unhealthy in blue line. We can see that most of unhealthy diet costs are

more expensive than healthy diet costs monthly for a college students, and unhealthy diet costs have a more violent fluctuation.

The last categorical bar plot shows the participants' monthly food consumption with the group of meal frequency "2 meals" or "3 meals". We can see that most of participants who take 2 meals cost less than participants who take 3 meals.

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

Methods

Confidence Interval

The confidence interval is the range of values that we expect your estimate to fall between a certain percentage of the time if we run the experiment again or re-sample the population in the same way.

I will invoke a non-parametric bootstrap [2] to derive the 95% confidence interval (CI) for the mean of food consumption per month on both healthy and unhealthy diet for college students.

$$\bar{x} \pm t_{n-1} \frac{s}{\sqrt{n}}$$

Hypothesis Test

Hypothesis testing in statistics is a way for you to test the results of a survey or experiment to see if you have meaningful results. You're basically testing whether your results are valid by figuring out the odds that your results have happened by chance. If your results may have happened by chance, the experiment won't be repeatable and so has little use.

The null hypothesis usually refers to the assumption of equality between population parameters. The alternative hypothesis is actually the opposite of the null hypothesis. Therefore, they are mutually exclusive and only one can be true. However, one of these two assumptions will always be true. So we can find out the validity of our hypothesis based on the result from hypothesis test.

We let the difference of mean of monthly food consumption on healthy diet for each college students and mean of monthly food consumption on unhealthy diet for each college be 50 dollars, which is a reasonable price for a motivation for college students to save money from healthy diet.

μ_1 = the mean of monthly food consumption on healthy diet for each college students.
μ_2 = the mean of monthly food consumption on unhealthy diet for each college. students.
Null hypothesis $H_0 : \mu_1 <= \mu_2$
The sample average college students food consumption is greater than or equal to the population average college students food consumption.
Alternative hypothesis $H_a : \mu > \mu_2$
We also set the alpha error $ equals to 0.05.

The one sample z-test we would like to perform is

$$\frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where n_1 = the sample size of monthly food consumption on healthy diet for college students.
n_2 = the sample size of monthly food consumption on healthy diet for college students.
$\bar{x_1}$ = the mean of monthly food consumption on healthy diet for college students in the sample.
$\bar{x_2}$ the mean of monthly food consumption on unhealthy diet for college students in the sample.
s_1 = the sample variance of monthly food consumption on healthy diet for college students.
s_2 = the sample variance of monthly food consumption on unhealthy diet for college students.

Results

Table 3: CI of food consumption on different dietary preference

	CI_healthy	CI_unhealthy
lower bound	556.5717	569.3722
upper bound	709.2177	748.4850

p-value
0.4022131

[3]There are 95% probability that we capture the true mean of food consumption per month on healthy diet for a college student from 556.5717 to 709.2177, and there are 95% probability that we capture the true mean of food consumption per month on unhealthy diet for a college student from 569.3722 to 748.4850.

The p-value is the probability that a given result would occur under the null hypothesis. The p-value is 0.4022131, which is greater than $\alpha = 0.05$. So we fail to reject H_0 that the sample average college students food consumption of healthy diet is greater than or equal to that of unhealthy diet. This is a reasonable result for having unhealthy diet cannot save money as the same conclusion we can make from the graph or the numerical summary.

Bibliography

1. College Student Eating Habits by Heidi Zwart | Company, higher education, Industry, wellness [https://www.betteryou.ai/college-student-eating-habits/]

2. Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: May 5, 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 surveyed:

```
## Rows: 47
## Columns: 9
## $ ID                     <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 13, 14, 15, ~
## $ gender                 <chr> "Male", "Male", "Male", "Male", "Female", "Male~
## $ monthly_living_expense <int> 1250, 1750, 1250, 750, 2250, 750, 1250, 1250, 7~
## $ proportions_on_food    <dbl> 0.3, 0.5, 0.3, 0.7, 0.3, 0.7, 0.5, 0.5, 0.7, 0.~
## $ dietary_preference     <chr> "Unhealthy", "Unhealthy", "Unhealthy", "Unhealt~
## $ meal_frequency         <chr> "2 meals", "2 meals", "2 meals", "3 meals", "3 ~
## $ snack                  <chr> "Maybe", "No", "No", "Maybe", "Maybe", "Maybe",~
## $ way_to_have_meal       <chr> "Dine in the restaurant", "Takeaway", "Home-mad~
## $ food_consumption       <dbl> 375, 875, 375, 525, 675, 525, 625, 625, 525, 52~
```