STA304 - Fall 2021

Assignment 1

Ruizhe Huang - 1006444331

Part 1

Goal

There are 1.4 million students at Canadian universities in 2019, and there are around 15% of them choose to live in the university's residence (Facts and stats, 2019). As Canada ranked 3rd globally in foreign student attraction in 2019 (El-Assal, 2020), thus the trend of university students would most likely increase. Then, we can infer that a fair amount of students will live in the university's residence in 2021. Moreover, it is undeniable that the dining hall food tastes are essential for the residents. Thus, the satisfaction of the residence's dining hall dishes is vital to help the university residence's manager improve the dining hall. This project is intended to survey the university residence's residents during 2021 in Canada. This project will investigate the residents' satisfaction with dishes in the university residence's dining hall. This topic is interesting and important since it will help the university's residence managers improve the dining hall's dishes.

Procedure

This project investigates the residents' satisfaction with university residence's dining halls. The target population will be all university residence's residents during 2021 in Canada. I need to contact with university in Canada first, tell them my research topic. Upon approval of the university, I strive for the residents' email addresses during 2021. Thus, the frame population will be all of the residents who had been registered in the university residence system during 2021. I will email them my survey, and then the sample population will be all of the residents who are still using the same email address as in the university residence system.
I will use the stratified sampling method by first group the residents by type of residents(international or domestic residents). And then randomly choose 1000 residents in each group. So, the sample will be 2000 residents who are still using the same email address as in the university residence system.
The strengths of the proposed sampling method are that since it is a random sampling, it is fairly representative and can make inferences about an overall target population. However, the drawbacks of the proposed sampling method can be nonresponse bias when I fail to elicit data from the residents since some of the residents may ignore my survey email. I will send the email again to try to reduce the nonresponse bias.
At last, the sample size would be 2000 university residences' residents who are still using the same email address as in the university residence system.

Showcasing the survey

The link of my survey: https://forms.gle/YYDbjfAeowmxAFEF8.

Question 1 How would you describe the university dining hall's dishes in 2021?
It is a linear scale question. The survey takers can scale the satisfaction of dining hall's dishes from 0(very unsatisfied) to 5(very satisfied). This question is one of the interested variables, and we will know

the residents' satisfaction with the dining hall's dishes with this scale. The benefit of this question is straightforward and consider the dishes taste may affect the satisfaction rate. The drawback of this question is that we can only get limited information. For example, we cannot get the reason for dissatisfaction.

Question 2 How many times approximately do you go to the university residence's dining hall per week in 2021?
It is a short answer question. The survey takers need to fill a whole number, the number of times they go to the dining hall per week. This question is one of the interested variables. We can know how often do the residents go to the dining hall each week. If they don't like the food in the dining hall, they may order delivery food instead of going to the dining hall. The benefit of this question is that we can analyze the results readily. The drawback of this question is that it needs the residents to recall the number of times they go to the dining hall per week, resulting in information errors if they forget.

Question 3 How long do you usually wait in the university dining hall each time in 2021? (in minutes)
It is a short answer question. The survey takers need to fill a whole number in this question, the usual wait time in minutes in the dining hall. This question is another interested variable. We will know the residents' average wait time in the dining hall. The benefit of this question is simple to answer and consider the wait time may affect satisfaction. The drawback of this question is that the residents need to recall the wait time, resulting in information errors if they forget.

Part 2

Data

Data simulation process

The dataset is a simulation. At first, we simulate the target population. I assume the number of all university residence's residents during 2021 in Canada is 200000.
I classified the residents into two types: international and domestic students with 25% of international residents and 75% of domestic residents with a total of 200000. Based on the study from Statista, there are 25% of international residents in the university in 2019. (Statista Research Department, 2020).
Then I classified whether the dining hall has weekly dishes with the answer 'Yes' or 'No' with the size of 200000.
Also, I categorized the customer service into 'Smiling', 'Poker-faced' and "Poor manners", with the size is 200000.
And then, I randomly select 200000 elements from the Normal distribution with the mean of 3 and the standard deviation of 5 to simulate the satisfaction rate of dining hall dishes in the population. The mean and standard deviation were determined based on the average satisfaction rate of dishes of university residents in Canada(Student Engagement Working Group Report, 2018). And, the range of satisfaction rate of dishes of university residents is not significant (Student Engagement Working Group Report, 2018). After that, I rounded all the decimals into the whole number as the satisfaction rate is a whole number. The Normal distribution was selected since the satisfaction rate is generally normally distributed (Usman Aziz, 2012).
Now I randomly select 200000 elements from the Poisson distribution with the mean is 15 to simulate the number of times residents go to the dining hall per week in the population. The Poisson distribution was selected since the number of times is generally a Poisson distribution, and it generates only whole numbers (Hayes, 2021). The mean was determined based on the average number of times university's residents go to the dining hall each day(more than 2 meals) times 7 days. (Student Engagement Working Group Report, 2018).
And I randomly select 200000 elements from the Normal distribution with the mean is 5 and the standard deviation is 10 to simulate the usual wait time in the dining hall in the population. The mean and standard deviation were determined based on the average wait time in the US university residence's dining hall. (Sara Brumfield, 2013), based on the range of wait time in the dining hall is relatively large (Sara Brumfield, 2013). After that, I rounded all the decimals into the whole number since most people only recall an approximate minutes of wait time, which is not accurate to seconds. The Normal distribution was selected since the wait time is generally normally distributed(Student Engagement Working Group Report, 2018).
Similarly, I randomly select 200000 elements from the Normal distribution with the mean is 30 and the standard deviation is 10 to simulate the usual eating time in the dining hall in the population. The mean and standard deviation were determined based on the average eating time in Canada's university residence's dining hall. (Student Engagement Working Group Report, 2018), based on the range of eating time in the dining hall is relatively large (Student Engagement Working Group Report, 2018). After that, I rounded all the decimals into the whole number since most people only recall an approximate minutes of eating time, which is not accurate to seconds. The Normal distribution was selected since eating time is generally normally distributed(Student Engagement Working Group Report, 2018).
Then I create the data frame called population contain these variables. Lastly, I randomly select 1000 residents from international and domestic students. In total, 2000 residents as the simulated dataset.
The dataset recorded the type of residents, customer service of university residence's dining hall, whether dining hall has weekly dishes, number of times residents go to the dining hall each week, average wait time in the dining hall and average eating time in the dining hall and the satisfaction rate of dining hall's dishes in 2021.
The original dataset contains 2000 observations and 8 variables. Each observation, in this case, represents a university residence's resident in Canada in 2021.

Data Cleaning Process

Since the original dataset is entirely from simulation and it doe not has any missing values. However, some negative numbers appear in the *satisfaction_rate*, *wait* and *eat_time* because they are randomly select from the normal distribution. Thus, I removed all the negative numbers as it makes nonsense in real life when the rate and time are negative.
Moreover, I created a new variable called the *satisfaction_of_dishes* classifies the satisfaction rate into two types: "Satisfied" or "Unsatisfied". Since the satisfaction_rate is the scale from 0-5, I let the rate greater than 3 be satisfied, and the others are unsatisfied.
Then, variables that were not useful was removed from the dataset, and only 7 critical variables were kept, including *residents*, *weekly_dishes*, *service*, *dining_freq*, *wait*, *eat_time*, and *satisfaction_of_dishes*. Since these variables are the main points for the assessment of university residence's dining halls.
For instance, *residents* classified the residents into international and domestic residents. Each type of resident may have different tastes. *weekly_dishes* provided whether this dining hall has weekly changed dishes so that residents could have more choice of food. *service* classified the staff attitude into smiling, poker-faced or poor manners. The service of staff directly affects the satisfaction of the dining hall. It's also interesting to know the number of times residents go to the dining hall per week. *dining_freq* could be an important variable related to the residents' satisfaction. If they go frequently, they may be probably satisfied with the dining hall's dishes. *wait* provided the usual wait time in the dining hall, we will know if residents wait a longer time, they probably less satisfied. *eat_time* provided the average time they spent eating in the dining hall. We will know if the residents eat more time, they are probably more satisfied with the food. *satisfaction_of_dishes* classified the university residences residents' satisfaction rate of dining hall's dishes into satisfied or unsatisfied.
The cleaned dataset contains 600 observations and 7 variables without any missing values.

Variable Description

Here are the important columns within the cleaned dataset.

Table 1: Description of important variables

Variables name	Description
satisfaction_of_dishes	Whether the university residences' residents are satisfied with the dining hall's dishes in 2021
dining_freq	The approximate number of times residents go to the university's dining hall per week in 2021
wait	The residents' usual wait time(in minutes) each time in the university's dining hall in 2021

Table 1 is a description of selected essential variables. I kept only a few variables to give readers a picture of the critical factors for each resident response. This report will focus on the following variables: *satisfaction_of_dishes*, *dining_freq* and *wait*.
satisfaction_of_dishes in the dataset representing whether the university residences' residents are satisfied with the dining hall's dishes in 2021. It is a categorical variable. It contains 'Satisfied' or 'Unsatisfied' as the response.
dining_freq in the dataset is the number of times residents go to the university's dining hall per week, approximately in 2021. It is a numerical variable. This variable was rounding into the whole number in the simulation process.
wait in the dataset is the residents' usual wait time(in minutes) each time in the university's dining hall in 2021. It is another numerical variable. This variable was rounding into the whole number in the simulation process as well.

Numerical summaries

This section will present the numerical summaries for the following two variables: the number of times residents go to the dining hall per week and the usual wait time in the dining hall each time.

Table 2: Numerical summaries of two important numerical variables

Variables	Mean	Median	Standard Deviation	Min	Max	IQR
number of times go to the dining hall per week	14.89	15	3.72	6	27	5
Usual wait time in the dining hall each time	9.79	8.5	7.07	0	41	10

Table 2 is the numerical summaries of the number of times residents go to the university residence's dining hall per week and the usual wait time(in minutes) in the dining hall each time in 2021. There will be some graphical summaries analyzing these variables later as well.
From Table 2, we know that among 600 residents, the sample mean and sample median of the number of times residents go to the dining hall per week is 14.89 and 15, respectively. So, It is slightly left-skewed. The sample standard deviation of the number of times residents go to the dining hall per week, which is 3.72 it expected a slight fluctuation of the data. The sample range of the number of times residents go to the dining hall per week is from 6 to 27, which is a slightly extensive range. The sample IQR of the number of times residents go to the dining hall per week is 5, which means that the range between the 1st and the 3rd quartile is 5, which is a relatively small spread.
For the usual wait time in the dining hall each time, the sample mean is 9.79 minutes, which is slightly higher than its sample median of 8.5 minutes. It implies that the usual wait time in the dining hall each time is right-skewed. The sample standard deviation of the usual wait time in the dining hall each time is slight significant at 7.07 minutes. It would expect a relatively large fluctuation of the data. The sample range is between 0 and 41 minutes, which is a fairly large range. The sample IQR is 10 minutes, which means that the range between the 1st and the 3rd quartile is 10 minutes, which is a relatively large spread. In summary, the usual wait time in the dining hall each time has a relatively more considerable fluctuation and spread.

Table 3 shows the proportion table of satisfaction of dishes in the dining hall (From 600 observations).

Table 3: Proportion of satisfaction of dishes in the dining hall

Satisfied	Unsatisfied
31.17%	68.83%

Table 3 is a numerical summary of the proportion of satisfaction of dishes in the dining hall. The Unsatisfied residents account for the most significant proportion, more than two times of the satisfied residents. There is also a graphical analysis of the proportion in the graphical summary, visualize the proportion of satisfied and unsatisfied residents. Furthermore, the satisfaction of dishes in the dining hall proportion will also be used for the bootstrap sampling in the results section.

Graphical summaries

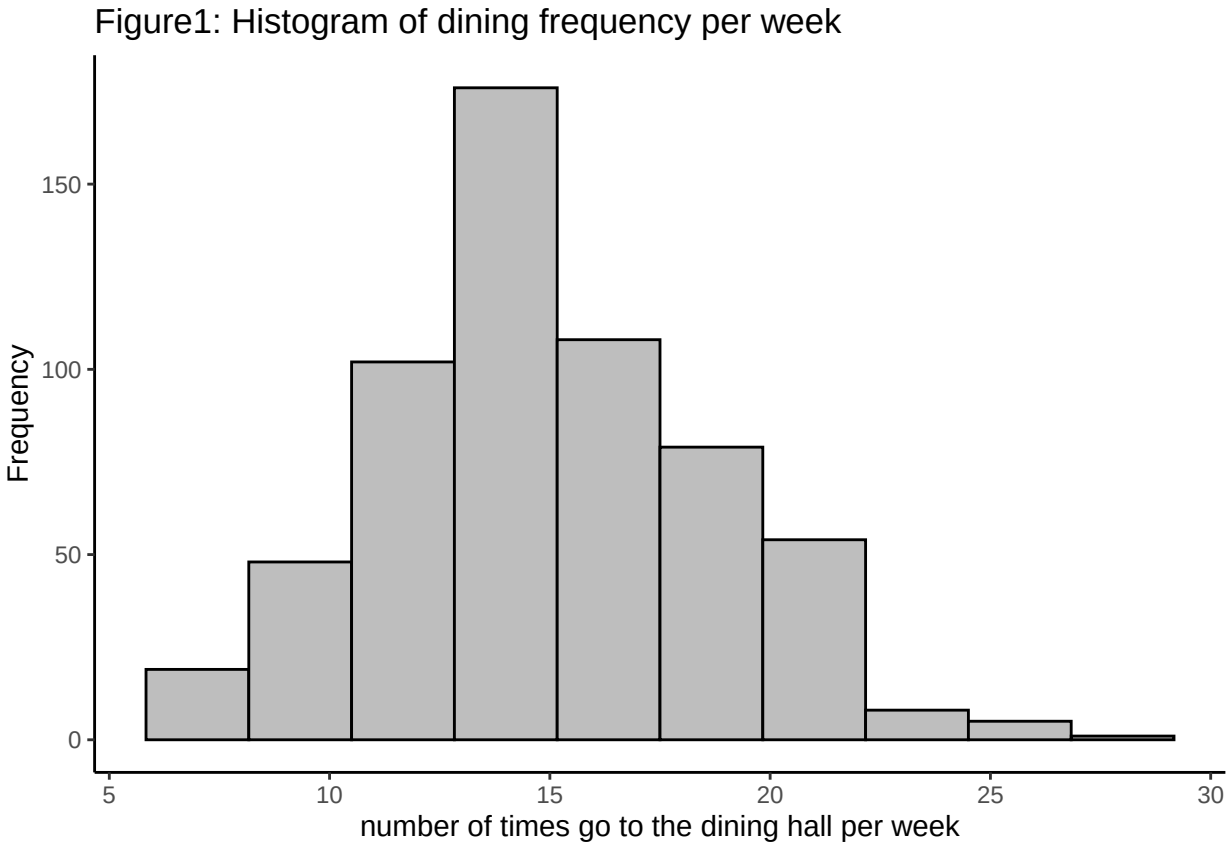

Figure 1 is the histogram of the number of times residents go to the university residence's dining hall per week in 2021. The center of the histogram is around 15. It means that most of the time, the residents go to the dining hall per week is approximately 15 times. The shape of this histogram is slightly right-skewed with one significant mode at the middle of the plot, which is around 15. From this figure, we found that the number of times residents go to the dining hall per week ranges between 5~30. It coincides with the range in the numerical summaries table, where the range is between 6~27 with a slight difference.
Moreover, most of the time, residents go to the dining hall per week is between 10~20 times. We cannot see any outliers in this plot. Therefore, the universities residences' residents go to the dining hall is around 15 times per week. According to this, most residents go to the dining hall frequently. That is more than two times in one day. And it is reasonable as we need to have three meals a day generally.

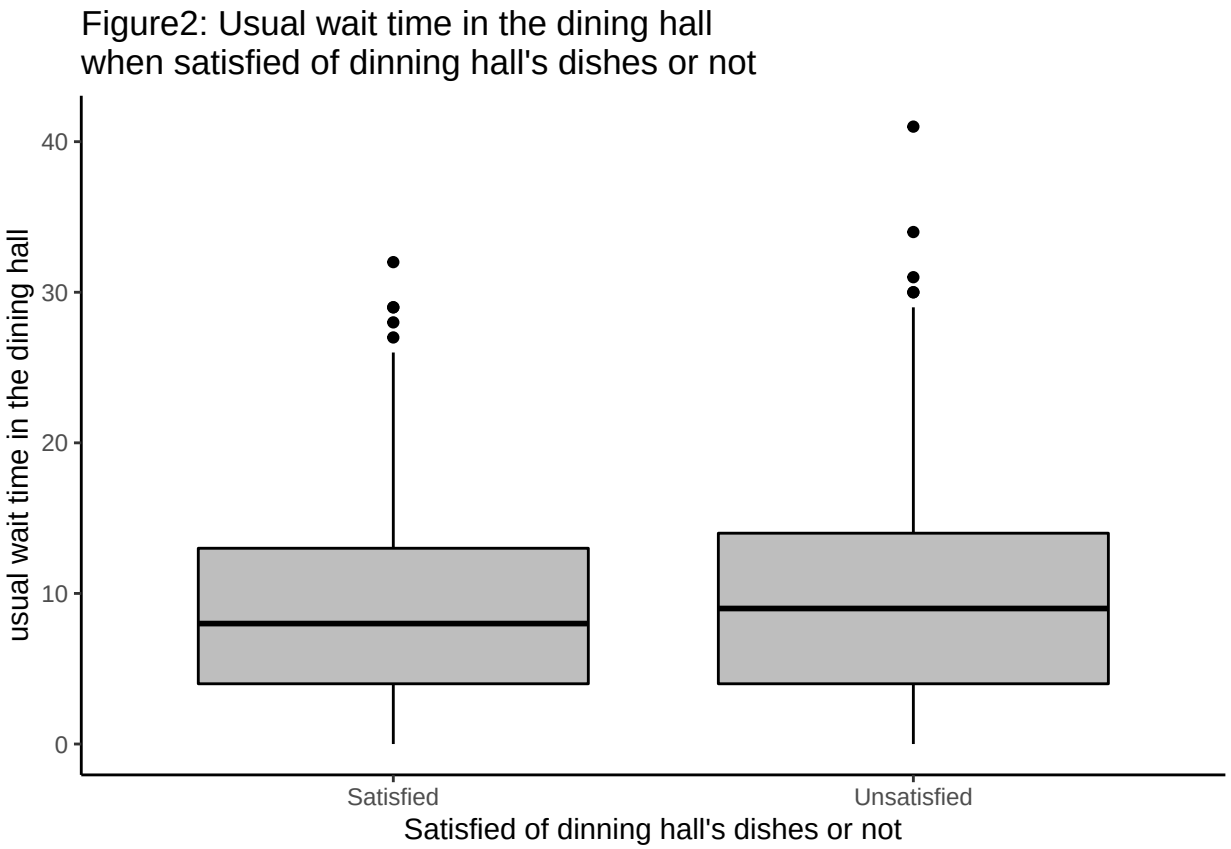

Figure 2 are the boxplots of the usual wait time in the dining hall if it has weekly dishes or not in 2021. For the residents satisfied with the dining hall's dishes in 2021, the boxplot is right-skewed with 4 outliers above 25 minutes. The median of the usual wait time in the dining hall is centred around 8 minutes. The first quantile nearly at 4 minutes. And the third quantile nearly at 15 minutes. So the IQR of the usual wait time in the dining hall is approximately 11 minutes. It implies that the spread of the usual wait time in the dining hall is moderate.

For the residents unsatisfied with the dining hall's dishes in 2021, the boxplot is right-skewed with 4 outliers above 28 minutes. The median of the usual wait time in the dining hall is centred around 9 minutes. The first quantile nearly at 4 minutes. And the third quantile nearly at 14 minutes. So the IQR of the usual wait time in the dining hall is approximately 10 minutes. It implies that the spread of the usual wait time in the dining hall is moderate.

In conclusion, regardless of the satisfaction of the dining hall's dishes, the residents are not that care about the wait time around 9 minutes as the satisfied and unsatisfied residents both wait for nearly 9 minutes. Moreover, the spread of satisfaction of dining hall's dishes is almost the same as unsatisfied. Therefore, the universities residences' dining hall's managers do not necessarily need to make too much effort on the wait time in the dining hall now.

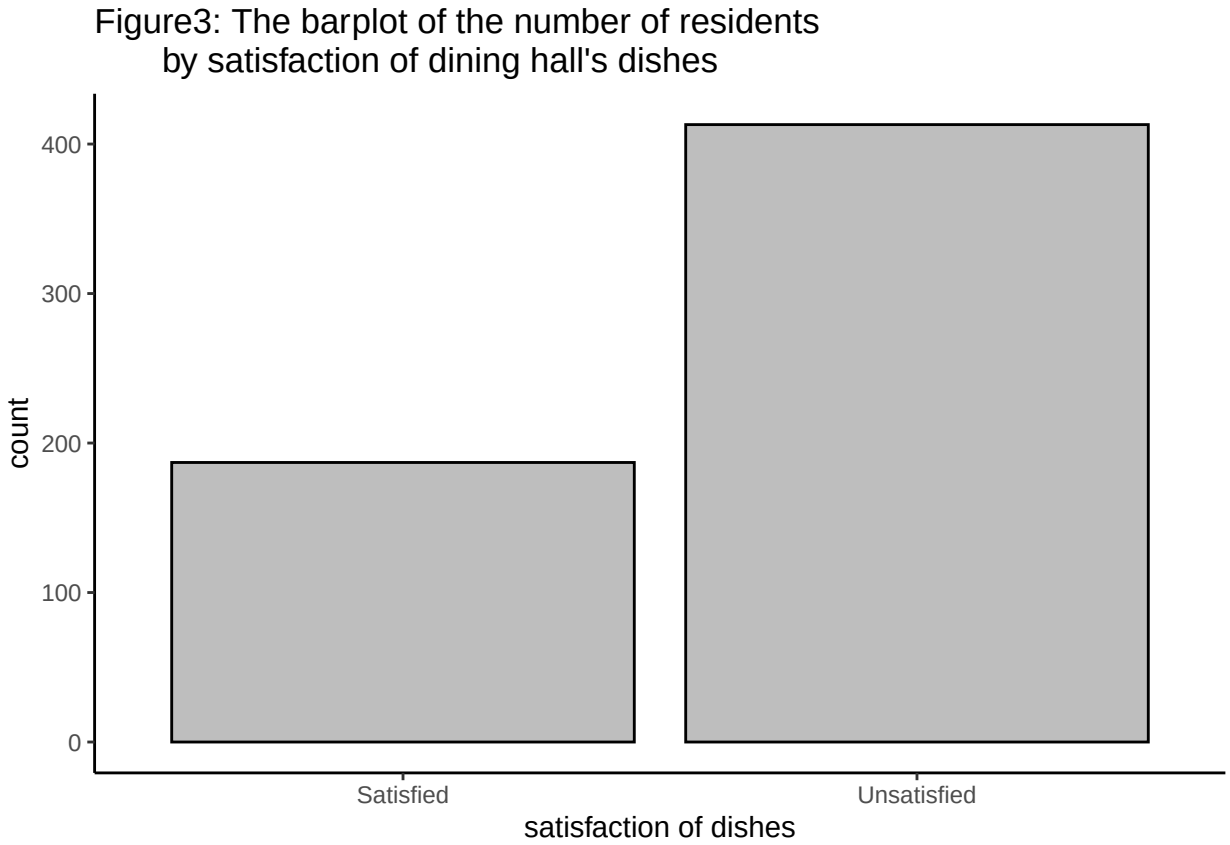

Figure 3 is a barplot of the number of residents by the satisfaction of dining hall's dishes(categorical variable), which were categorized by satisfied and unsatisfied. The unsatisfied residents accounted for the most significant part, and it's more than two times of the satisfied residents. There are 600 university residences' residents. Nearly 190 were satisfied with the dishes, and 410 were unsatisfied with the dishes in the dining hall.
Figure 3 shows that more than 60% of people are unsatisfied with the dining hall dishes in the university residence. Thus, the dining hall's managers could improve the food to make more people satisfied.
Furthermore, I will use the satisfied proportion to perform the bootstrap sampling with confidence interval in the next step. Specifically, I want to look into what will the true proportion be like after a thousand simulations. Thus, we will be using the proportion of the satisfied residents to do the bootstrap confidence interval in the results section.

Methods

Assume the data points in the sample are independently and identically distributed.

Hypothesis Test

As we have a large sample size, which contains 600 observations. By Central Limit Theorem(CLT), then the sample mean would approximately follow the normal distribution. I will construct the hypothesis test of the mean for the usual wait time in the dining hall in 2021 by the following steps to see whether the observed data support my hypothesis or not.
Step1: Formulate a null (H_0)and alternative hypothesis (H_A).
Let μ be the population mean of the usual wait time in the dining hall in 2021.

Since the usual wait time in the dining hall in 2021 is around 9 minutes.
Therefore, I hypothesize that:
$H_0 : \mu = 9$
$H_A : \mu > 9$
Step2: Find the test statistic. Since I have a relatively large sample size in the dataset, so the T-test stat will asymptotically approach a Z-test stat. Then, I will use T-test get the test statistic(t).

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$$

where $\bar{x}$ is the sample mean; μ_0 is 9 as the population mean under H_0 is true; s is the sample standard deviation; n is the sample size; t is the t distribution.
Step3: Calculate the corresponding p-value.
A p-value is a measure of the probability that an observed difference could have occurred just by random chance (Beers, 2021). The smaller the p-value, the stronger the evidence that we should reject the null hypothesisH_0.

$$\text{p-value } = P(|t_{n-1}| > |t|)$$

Step4: Make a decision based on the significance level (α)
I will pre-specified the significance level $\alpha = 0.05$ as a cut-off for rejecting or NOT rejecting the null Hypothesis H_0. If p-value < 0.05 then we reject H_0. If p-value > 0.05 then we do NOT reject H_0.
The variable *wait* is used in this method. It is a numerical variable.

Confidence Interval

Bootstrapping is used to estimate sampling distribution by re-sampling from the original sample with replacement in the same size. It allows us to explore the variability of estimates from the original sample.
I will invoke an empirical bootstrap to derive the 95% confidence interval for the population proportion of satisfaction rate of the dining hall's dishes in 2021. The empirical bootstrap sample is a re-sample where a number of smaller samples of the same size are repeatedly drawn, with replacement, from a single sample data to estimate bootstrap sampling distribution. (Stephanie, 2016)
For the assumption of empirical bootstrap, the sample data should be a representative random sample of the population, and the samples are independent identically distributed with some unknown distribution function. The cleaned dataset is simulated and includes 600 university residence's residents in 2021. We assume that this dataset contains random samples without much bias from the population.
Confidence interval is a range of values likely to include a population parameter of interest with a certain degree of confidence. Since I want to have higher confidence, but without the confidence interval is too broad. Therefore, I will construct 95% confidence intervals through empirical bootstrapping later in the result section. It means that I am 95% confident that the true population parameter is between the upper and lower bounds of the confidence interval. I will get the upper and lower bounds by calculated the 2.5th percentiles of bootstrapping simulation results as the lower bound and 97.5th percentiles of bootstrapping simulation results as the upper bound.
I will iterate the simulation 1000 times in the results section. After the simulation of empirical bootstrapping, I will get the confidence interval of the variable of interest, where the parameter of interest is the proportion of satisfaction rate of the dining hall's dishes in 2021.
The variable *satisfaction_of_dishes* is used in this method. It is a categorical variable.

Terminology

CLT: The central limit theorem states that as the increase of sample size, the distribution of sample variables approximates normal distribution, all samples are assumed to be the same size, regardless of the actual distribution shape of the population. (Ganti, 2021)

Results

Hypothesis Test

I have hypothesized in the method section that:
$H_0 : \mu = 9$
$H_A : \mu > 9$

Table 4: Mean of the usual wait time in the dining hall hypothesis test

Variable name	$\hat{\mu}$	significance level(α)	Test Statistics	p-value
usual wait time in the dining hall	9.79	0.05	2.7268	0.0033

Table 4 is the summary table of the result of the mean of the usual wait time in the dining hall hypothesis test. We have the sample mean usual wait time in the dining hall is 9.79 minutes since p-value < 0.05 then there is strong evidence to support H_A that is the mean usual wait time in the dining hall is more than 9 minutes. Based on the observed data, we do not have enough evidence to reject the hypothesis that the population mean usual wait time in the dining hall is more than 9 minutes. The result seems reasonable since the sample mean is 9.79 minutes, which is more than 9 minutes. Also, it is practically common that the mean usual wait time in the dining hall is more than 9 minutes. Therefore, the university residences' dining halls would expect a longer wait time over 9 minutes generally.

Confidence Interval

Assume the sample data is a representative random sample of the population, and the samples are independent identically distributed with the Normal distribution. Table 5 is a 95% confidence interval for the proportion of satisfaction rate of the dining hall's dishes in 2021.

Table 5: 95% Confidence interval of satisfaction rate of the dining hall's dishes

Parameter	Confidence level	Confidence interval
Satisfaction rate of the dining hall's dishes	95%	[27.5%, 35%]

From table 5, the 95% confidence interval for the population proportion of satisfaction rate of the dining hall's dishes is [27.5%, 35%]. It means that we are 95% confident that the true proportion of satisfaction rate of the dining hall's dishes is between 27.5% and 35%. The result seems reasonable since it's close to the sample proportion of satisfaction rate of the dining hall's dishes. Moreover, it is practically common that most residents from foreign countries or immigrate from other countries are not used to Canadian food. Therefore, the university residences' managers should be aware that only 27.5% to 35% of residents were satisfied with the dining hall's dishes. The dining halls' managers could make some improvements to the dishes.

Bibliography

All analysis for this report was programmed using `R version 4.0.3`.
1. Allaire, J.J., et. el. *References: Introduction to R Markdown*. RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: September 30, 2021)
2. Beers, B. (2021, September 28). *What P-value tells us. Investopedia*. Retrieved October 1, 2021, from https://www.investopedia.com/terms/p/p-value.asp.
3. Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how*. Springer Science & Business Media.
4. Distribution of Customer Satisfaction Survey. iSixSigma. (n.d.). Retrieved October 1, 2021, from https://www.isixsigma.com/topic/distribution-of-customer-satisfaction-survey/.
5. El-Assal, K. (2020, February 20). *642,000 international students: Canada now ranks 3rd globally in foreign student attraction*. CIC News. Retrieved October 1, 2021, from https://www.cicnews.com/2020/02/642000-international-students-canada-now-ranks-3rd-globally-in-foreign-student-attraction-0213763.html#gs.cqeqqs.
6. Facts and stats. *Universities Canada*. (2019, August 31). Retrieved September 29, 2021, from https://www.univcan.ca/universities/facts-and-stats/.
7. Ganti, A. (2021, September 21). *What is the Central Limit Theorem (CLT)? Investopedia*. Retrieved September 30, 2021, from https://www.investopedia.com/terms/c/central_limit_theorem.asp.
8. Grolemund, G. (2014, July 16) *Introduction to R Markdown*. RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: September 30, 2021)
9. Hayes, A. (2021, October 1). Poisson distribution. Investopedia. Retrieved October 1, 2021, from https://www.investopedia.com/terms/p/poisson-distribution.asp.
10. journalist, S. B. M. staff. (2013, September 1). Students not impressed with long lines at dining halls. The Athens Messenger. Retrieved October 1, 2021, from https://www.athensmessenger.com/news/students-not-impressed-with-long-lines-at-dining-halls/article_a5c2427e-3da2-5fe2-b03f-83bba9cfe6c3.html.
11. Statista. *Accommodation type of international students in Canada 2018*. (2020, November 6). Retrieved September 29, 2021, from https://www.statista.com/statistics/946404/international-students-housing-canada/.
12. Stephanie. (2016, November 03). *Bootstrap sample: Definition, example*. Retrieved September 29, 2021, from https://www.statisticshowto.com/bootstrap-sample/
13. Student Engagement Working Group Report meal ... - ualberta.ca. (n.d.). Retrieved October 1, 2021, from https://www.ualberta.ca/residence/media-library/documents/student-engagement-working-group-report-meal-plan-consulation.pdf.

Appendix

Here is a glimpse of the original dataset of simulation:

```
## Rows: 2,000
## Columns: 8
## $ id                <int> 174519, 86323, 146925, 163011, 46413, 2066, 190830, ~
## $ residents         <chr> "domestic", "domestic", "domestic", "domestic", "dom~
## $ weekly_dishes     <chr> "Yes", "Yes", "No", "No", "Yes", "No", "No", "Yes", ~
## $ service           <chr> "Poker-faced", "Poor manners", "Poker-faced", "Poor ~
## $ satisfaction_rate <dbl> -3, -1, 4, 5, 1, 6, -4, 1, -11, 10, 4, 5, 2, -3, -1,~
## $ dining_freq       <int> 11, 17, 13, 13, 16, 12, 14, 10, 16, 18, 15, 9, 15, 1~
## $ wait              <dbl> 6, 15, 25, 13, 11, 0, 17, 7, 11, -2, 6, -6, -5, 12, ~
## $ eat_time          <dbl> 31, 25, 23, 33, 22, 19, 26, 26, 34, 23, 35, 19, 38, ~
```

Here is a glimpse of the cleaned dataset of simulation:

```
## Rows: 600
## Columns: 7
## $ residents              <chr> "domestic", "domestic", "domestic", "domestic",~
## $ weekly_dishes          <chr> "No", "No", "Yes", "Yes", "No", "No", "Yes", "N~
## $ service                <chr> "Poker-faced", "Poor manners", "Smiling", "Smil~
## $ dining_freq            <int> 13, 13, 16, 10, 15, 10, 8, 8, 20, 19, 11, 18, 1~
## $ wait                   <dbl> 25, 13, 11, 7, 6, 8, 12, 2, 17, 15, 7, 11, 7, 1~
## $ eat_time               <dbl> 23, 33, 22, 26, 35, 45, 28, 21, 40, 25, 48, 23,~
## $ satisfaction_of_dishes <chr> "Satisfied", "Satisfied", "Unsatisfied", "Unsat~
```