# Connection between Residential Choices and Campus Life

## Statistics 302 Accelerated Introduction to Statistical Methods

*Team 16: Panda*

*Yuqi Niu, Jing Duan, Yanxu Guo, Zimo Wan*

*Dec. 20th, 2019*

# Contents

# 1. Abstract

The goal of the study is to find the connection between residential choice and campus life for Chinese students who studying at the University of Wisconsin Madison. The population of interest is all current Chinese students at the University of Wisconsin Madison. A survey is popularized through the Wechat app, a famous Chinese Social media, to gather data. The case of the study is 70 Chinese students who answered the survey and are stratified by gender from the original data. The cases are selected randomly to avoid sampling bias. Moreover, the relationships between variables are tested in 4 types of statistical test, such as randomization test, t-test, chi-square test, and linear regression. 5 variables are tested in the study. The explanatory variables that are tested are students' on-campus job status, the cost of housing, and students' housing choice. The response variables are whether always cook by themselves, the students' housing choice, and the number of times students eat in the dining hall. The t test shows the cost of apartment is lower than the cost of the residence hall, and the chi-square test reveals that most students live in apartments like to cook by themselves, which further indicates cost and availability of cooking are factors that attract students to live in apartments. However, the randomization test shows that there is no difference between the frequency of eating in the dining hall per week for students with and without on campus jobs, and the regression test shows there is no relationship between times of eating in the dining halls and cost of housing. The connections and disconnections resulted from the study both provided useful guides for future Chinese students to make residential decisions and for school's service institutions to improve their services.

# 2. Introduction

## 2.1 Background Research

The University Housing website[4] shows that there are 7800 students living on campus each year at the University of Wisconsin Madison. Various factors influence students' choice of living on campus or in apartments. The low residential fee is one factor that attracts students to live on campus. The University of wisconsin Madison offers the lowest on-campus housing price in the big ten. Besides, the grade of students is another important factor. Each year, over 90 percent of freshmen live on campus to enrich their Wisconsin Experience, which contributes to the majority of campus housing residents, and many of them choose to return during their sophomore year. It is interesting to test what are the factors influencing students' choice.

Similar research has been conducted on the factors influencing students' housing choice. According to the research conducted by Oyetunji Abiodun Kolawole and Abidoye Rotimi Boluwatife (2016)[1] in Nigeria, the housing costs and accessibility to school buildings are two critical factors that influence students choice on housing most. The cases in this study are students from a university in Nigeria, which are at similar ages as the cases in the designed study. Another research was conducted by Kobue, Oke, and Aigbavboa (2017)[2] in South Africa, showing that students choose where to live based on various factors, including the location of residence, proximity to campus, and whether the building contains different kinds of study rooms. In both of the studies, factors, such as distance from campus and cost of housing, play an important role in student decision of housing. Since different countries may have cultural variations, the factors influencing students of other countries is different from students in the University of Wisconsin. Therefore, it is necessary to investigate what is the localized factors for Chinese students in UW-Madison.

Moreover, "Away from Home: Chinese International Students at UW-Madison"[3] reveals that there were about 30,000 undergraduate students at UW-Madison in 2017. Among all undergraduate students, 8% were international students, with 1,661 Chinese students, or about 52% of all international undergraduate students.This data shows that Chinese undergraduate students are in great amount and the data gathered from them may be representative enough. The focus of this study is on a Chinese students in UW-Madison and the results can be utilized in practice within the school while previous studies were focusing on the housing choice of students in other countries, which is not applicable to chinese students at UW-Madison. Therefore, the population of interest in this study is different from previous studies.

## 2.2 Impact and Significance of the Study

The results generated from the study could give incoming or current Chinese students an effective guide on where to live based on their anticipated housing price and live-styles. This is crucial for Chinese freshmen who leave their home for the first time and enter a new environment. The experience from the former Chinese students could be a useful reference for their housing Choice.

In school housing service's perspective, the result of this study could be used for the residence halls at UW-Madison to improve its marketing strategies to attract more students. For instance, if there is a strong indication that students living in the apartment paylower price, which ends up more students living in apartment, then the university may think of lower its housing price to attract more students and gain more profit. The study may also promote the improvement of residential devices. For instance, if most Chinese students who cook by themselves live in apartments, students may like to cook, and university residence halls should offer more kitchen in order to attract students who like to cook.

## 2.3 Statement of Purpose

The purpose of the study is to determine the factors that influence Chinese students' housing choice and the influence of housing decision on students residential life. Generally, the study is designed to determine

how the aspects of students college life, which include job and demographics, influencing Chinese students' housing choices, such as either live in residence hall or apartment. The assumptions are Chinese students are inclined to choose house that is cost-efficient and achieve their live preference. The result of the study could further be used to guide incoming students' housing choices and the policy making of the school.

The aim of this study is to investigate what factors influence the choice of accommodation for Chinese students at UW-Madison and to what extent does each variable influence student choice. Specifically, those factors include variables such as the grade the students, having on-campus jobs or not, and the costs of housing, and whether students cook by themselves. In addition, the effect the accommodation choice has on Chinese students' frequency of eating in the dining hall, the frequency of cooking was also investigated in the study. In order to achieve the purpose of this study, an online surveys are sent to Chinese students at UW-Madison.

# 3. Methods

## 3.1 Data Collection

### 3.1.1 Population of Interest

The population of the study is all Chinese students who are currently studying at UW-Madison. Although Chinese student may not represent the choice of local white students or other international students, the data is valuable for school's dining and housing service policy making and Chinese students' dining and housing decision making.

### 3.1.2 Type of Study

The study is an observational study. Since it is not feasible to assign participants to live in a particular dorm or behave in a certain way, the researchers choose to record observations and only gain information from the answers of the participants in the questionnaires, without making any control. The analysis of the study was based on the fact of the factors influence students' choice instead of the result of randomized experiment.

### 3.1.3 Survey Protocol

The location of the survey is online through Wechat, and it was sent out on Nov.18th. Since the peak time for students to check social media is about 8 or 9 pm, the survey was posted mainly during that time. In addition, the survey was sent out by all four members of the group to collect more answers to increase the dataset to make it more representative.

The data was collected using an online questionnaire system called Google Forms to design an anonymous well-worded questionnaire concerning those 8 variables and posting it on WeChat (a popular Chinese social media), and those who see the questionnaires may answer it. After collecting data, if the sample size is not large enough to use statistical test, then population parameters are estimated through bootstrap method, the spread and the center is indicated through ggplot to indicate the relationships between variables, and the data was further analyzed through randomization test, t test and Chi-square test.

### 3.1.4 Randomization Protocol

Firstly, the survey was randomly send out in different chat groups, including renting group, second-hand items selling group, food ordering groups, and general students' groups organized by Chinese student association. Those groups contain chinese students with different grades and with different gender.

Therefore, the students who answered the surveys are likely to be random students who are not the friends of the researcher.

Secondly, the 70 samples of this project were picked randomly from the original 111 samples. Considering the ratio of female students and male students at University of Wisconsin-Madison is 1:1, the researcher stratified the sample by randomly choosing 35 males and 35 fales.

The economic condition could be one of the confounding variables in the research. The wealthier a student is, less likely he or she will have campus job, and the fewer times he or she will eat in dining halls. Since this could have an impact on both the explanatory and response variables, it is something we might have to consider.

#### 3.1.5 Randomness and Representativeness

The data of this study that gained from the survey may not be random and representative, but through the stratified sampling based on the data collected, the sample can become more representative.

Possible bias may include convenience sampling bias. Since it is not feasible to reach out to every Chinese student at UW-Madison, the survey may only be spread to students who have mostly frequent contacts with researchers.

To reduce the bias above, stratified sampling regarding the variety of the targets was used. The online questionnaire was posted in different chat groups (such as renting groups, study groups), which are not limited to the friend groups of researchers. Since the Chinese social media platform WeChat is prevalent among Chinese students, the questionnaire was likely to be seen by a large number of Chinese individuals with different genders and other characteristics. The diversity of the participants is ensured.Therefore, the sample is likely to be relatively representative of the whole population.

Another potential bias is non-response bias. Since students are not forced to participate in this survey, the majority of them might just ignore the questionnaire, which may not only cause the sample size to be too small to analyze, but it may also decrease the representativeness of the samples to the population.

Some amount of reward for students who fill the survey could be used to avoid the situation above. Small amount of money is sent out along with questionnaire in group chats on WeChat. Though the money amount might be just several bucks, the response rate of students might increase significantly.

#### 3.1.6 Sample Size

In order to use Central Limit Theorem to approximate normal distribution, the sample size should be at least 30. The sample size of the study is 70, consisted of 70 Chinese students who answered online questionnaires, including 30 students living in residence halls and 40 students living in apartments. The sample is stratified by gender from 111 students who answer the survey. Since the ratio of female students and male students in the university of wisconsin madison is 1:1, the sample is stratified with the ratio with 35 males and 35 fales. Therefore, this sampling method tends to produce a good representative sample of all chinese students in the university of wisconsin madison.

- For the cost of housing, based on the 111 survey responses collected, the standard deviation is 250.46 and the t statistic is -2.14. When the margin of error is 65 dollars, the sample size needed is

$$\begin{aligned} n &= (\frac{t * SE}{ME})^2 \\ &= (\frac{-2.14 * 250.46}{65})^2 \\ &= 67.99 \end{aligned}$$

  Since the sample size used in this study is 70, which is more than the sample size of 68 needed, together with each category is at least 30, the sample collected is good for t test to test the mean monthly cost of housing with margin of error equals to 65 dollars.

## 3.2 Variables of Interest

| Name of Variable (Shorthand Notation) | Type of Variable | Level of Variable (If Categorical) |
|---|---|---|
| Grade (Grade) | Categorical Explanatory | A: Freshman, B: Sophomore, C: Junior, D: Senior, E: Super Senior/Graduate |
| Job (Whether have a job) | Categorical Explanatory | A: Yes, B: No |
| OnCamp (Whether the job is on campus) | Categorical Explanatory | A: Yes, B: No |
| Time (Time on foot from your living place to your major building that you go to most frequently) | Categorical Explanatory | A: 5-10 minutes, B: 10-15 minutes, C: 15-20 minutes, D: more than 20 minutes |
| Cost (Cost of housing) | Numerical Explanatory | |
| Credits (Credits taking this semester) | Numerical Explanatory | |
| Housing (Living in residence hall or apartment) | Categorical Explanatory & Response | A: Residence hall, B: Apartment |
| Cooking (Do you usually cook by yourself) | Categorical Response | A: Yes, B: No |
| Roommate (Number of roomate you have) | Numerical Response | |
| Dining (Number of times you go to the dining hall per week) | Numerical Response | |
| Party (Number of times you attend a party per week) | Numerical Response | |

## 3.3 Statistical Analysis

### 3.3.1 Type of Statistical Tests

This study used statistical tests such as randomization test, z-test, t-test, and chi-square test to analyze the relationships between the variables. Null hypothesises and Alternative hypothesises for the test are listed as below.

1. Randomization Test:
   $\mu_1$=the average number of times students who have on campus jobs eat in the dining hall per week
   $\mu_2$=the average number of times students who do not have on campus jobs eat in the dining hall per week
   $$H_0 : \mu_1 = \mu_2$$
   $$H_a : \mu_1 \neq \mu_2$$
2. T Test:
   $\mu_1$ = population mean cost of apartment
   $\mu_2$ = population mean cost of residence hall
   $$H_0 : \mu_1 = \mu_2$$
   $$H_a : \mu_1 < \mu_2$$
3. Chi-square Test:
   $$H_0 : Students' \ cooking \ frequency \ is \ not \ associated \ with \ their \ housing \ type$$
   $$H_a : Students' \ cooking \ frequency \ is \ associated \ with \ their \ housing \ type$$
4. Linear Regression:
   Linear Regression: $\beta$ is the slope of the least squares line to predict the cost of housing based on the times students eat in the dining hall.
   $$H_0 : \beta = 0$$
   $$H_a : \beta \neq 0$$

#### 3.3.2 Computational Requirements

This study uses R Studio as the computational software to do the statistical tests on the data and to graph the distributions. The packages used in the process of analysis are tidyverse, dplyr, ggplot2, grid, gridExtra, knitr, reshape2, ggfortify, xtable.

# 4. Results

## 4.1 Randomization Test: Campus Job and Student Dining Choice

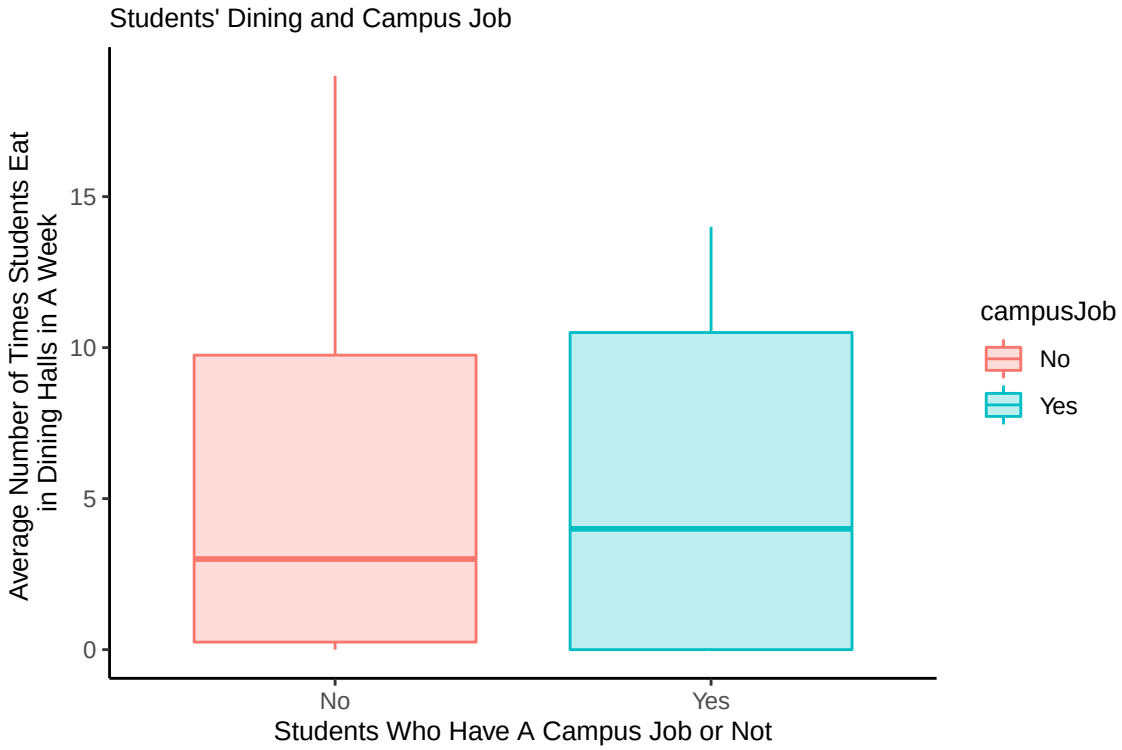
The variables of the randomization test are mean number of times students eat in the dining hall per week, which is quantitative variable that measures how often students eat in the dining hall every week, and employment status, which is divided into students with on-campus job and students without on-campus job. A question derived from the research problem of this project is whether having on-campus job influence the average number of times students eat in the dining hall per week. The hypothesis of this test is that there is a difference between the average number of times students eat in the dining hall per week for students with and without on-campus jobs. Therefore, the purpose of this test is to examine if the average number of times students who have on-campus job eat in the dining hall per week is different from the average number of times students who don't have on-campus job eat in the dining hall per week.

#### Hypothesis

$\mu_1$=the average number of times students who live in dorms eat in the dining hall per week
$\mu_2$=the average number of times students who live in apartments eat in the dining hall per week

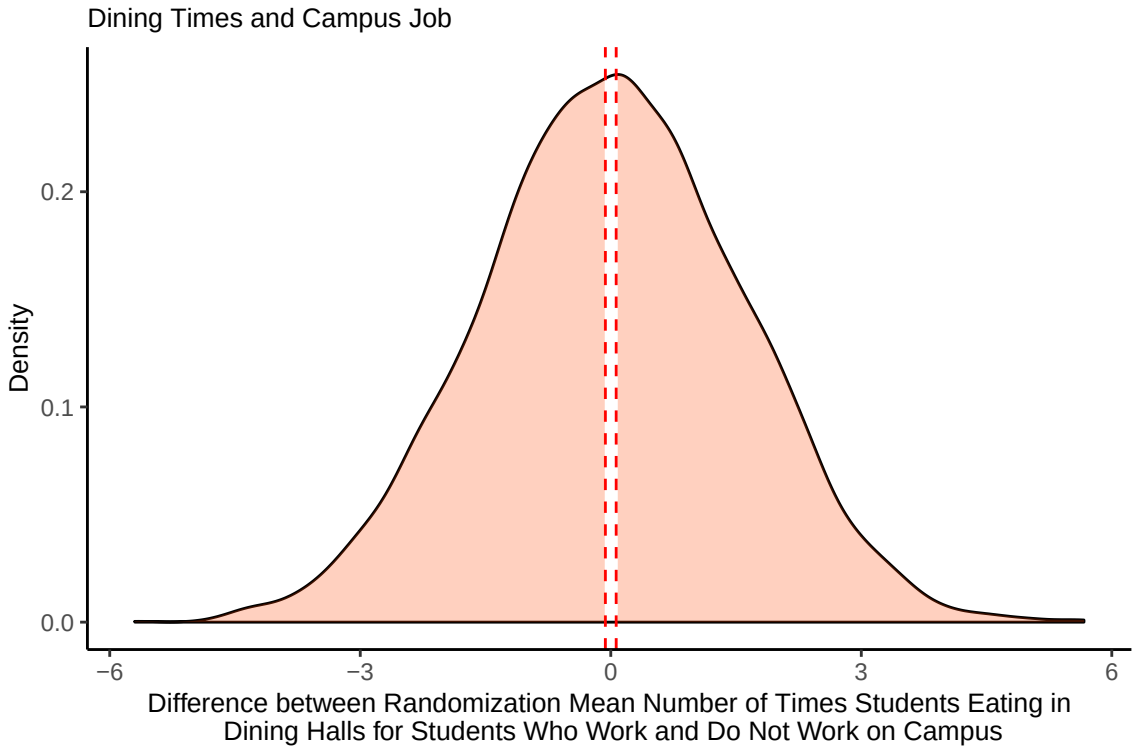
$$H_0 : \mu_1 = \mu_2$$
$$H_a : \mu_1 \neq \mu_2$$

**Summary Figure**




Since the spreads of the two groups are somewhat different, we generated two randomization distributions separately assuming their means are equal and then take the difference.

**Checking Assumption**




The randomization distribution is approximately symmetric and bell-shaped, so randomization test could be used. The dash lines represents the sample mean differences(positive and negative), while the shaded area represents the possibility that a larger difference exists.

**Test Statistics**

Difference of mean number of times eating in dining halls between groups of students who has and has no campus job:
$\mu_1$-$\mu_2$=0.06481481

**Compute p-value**

Define rand.diff as the difference of the sample means of number of times eating in dining halls between groups of students who has and has no campus job generated by randomization.

$$P - Value = P(|rand.diff| \geq |\mu_1 - \mu_2|) = P(|rand.diff| \geq 0.06481481) = 0.9663$$

The p-value calculated based on the assumption that the null hypothesis is true (the difference between number of times students who have and have no campus jobs eat in dining halls is zero)is 0.9663. The shading area in the density plot above also illustrates the large p-value.

**Confidence Interval**

The 95% confidence interval for the difference in the mean number of times students eat in the dining hall per week between students with and without on campus jobs is (-3.041667 ,3.023380).

**Interpretation**

Since p-value of the randomization test is 0.9663, there isn't significant evidence that there are differences between the average number of times students eat in the dining hall per week between students with and without on-campus jobs. (randomization test for difference in mean, $\mu_1$-$\mu_2$=0.06481481, p=0.9663, $\alpha$=0.05)

## 4.2 T-Test: Housing Choice and Cost of Housing

The variables of this test are cost, a quantitative variable that measures how much students pay for their living place each month, and the choice of housing, a categorical variable that is divided into apartment and residence hall. Another question derived from the research question is whether a cheaper cost is associated with more students choose to live in the place. The hypothesis in this project is that apartments have smaller mean cost than residence halls. Hence, the aim of this test is to test if the mean cost of the apartment is lower than that of the residence house.

**Hypotheses**

Define $\mu_a$ as the population mean monthly cost of living in an apartment and $\mu_r$ as the population mean monthly cost of living in a residence hall.

$$H_0 : \mu_a = \mu_r$$
$$H_a : \mu_a < \mu_r$$

**Summary Figure**

**Check for Assumptions**

| Housing | Count |
|---|---|
| Apartment | 40 |
| Residence Hall | 30 |

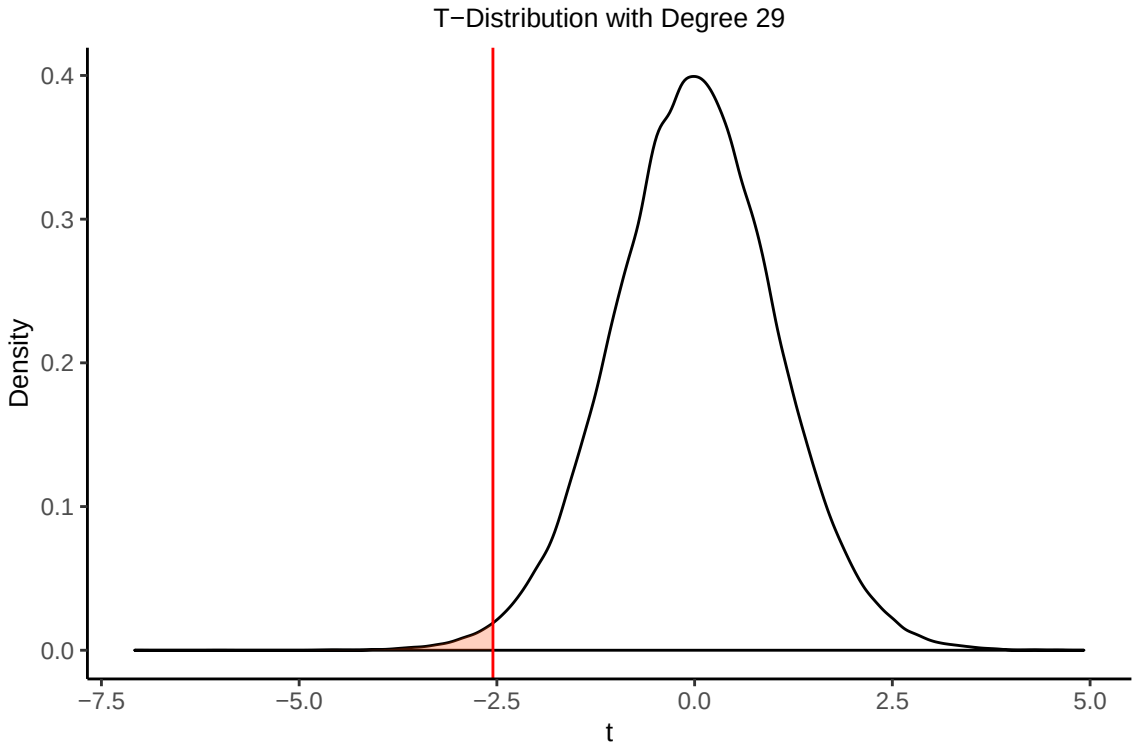
The underlying distribution is approximately normal distribution, and the sample size is large enough with $n_a \geq 30$ and $n_r \geq 30$ where $n_a$=40 and $n_r$=30. Therefore, a t-distribution is appropriate.

**Calculate Test Statistic**

Define $\bar{x}_a$ as the sample mean cost of apartments, $\bar{x}_r$ as the sample mean cost of residence halls, $s_a$ as the standard deviation of the sample costs of apartments, $s_r$ as the standard deviation of the sample costs of residence halls, $n_a$ as the number of observations of students who living in apartments, and $n_r$ as the number of observations of students who living in residence halls.

$$\begin{aligned} t &= \frac{\bar{x}_a - \bar{x}_r}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_r^2}{n_r}}} \\ &= \frac{927.500 - 1040.833}{\sqrt{\frac{227.852^2}{40} + \frac{142.406^2}{30}}} \\ &= -2.551 \end{aligned}$$

**Compute P-Value**



$$P - Value = P(T \leq t) = P(T \leq -2.551) = 0.00814$$

This probability is corresponding to the area to the left of the red line in the t-distribution plot.

#### 95% Confidence Interval

The 95% confidence interval for the mean monthly cost of living in an apartment is 24.63 to 202.04 dollars lower than that of living in a residence hall.

#### Interpretation

There is significant evidence that the mean cost of apartment is lower than the mean cost of residence hall. (lower tiles independent t-test, $\bar{cost}_a = 927.5$, $\bar{cost}_r = 1040.833$, $n_a = 40$, $n_r = 30$, $t = -2.5509$, $df = 29$, $p = 0.008141379$, $\alpha = 0.05$).

## 4.3 Chi-square Test: Cooking Frequency and Housing Choice

The variables of the chi-square test are housing choice, which is a categorical variable that is divided into residence hall and apartment, and cooking frequency, which is the other categorical variable that is divided into often cook and not often cook. A question derived from the research problem of this project is whether living in apartments or residence halls is associated with how often students cook. The hypothesis of this test is that there is no association between the housing choice and students' cooking frequency. Therefore, the purpose of this test is to examine if students' cooking frequency is associated with where they live, residence halls or apartments.

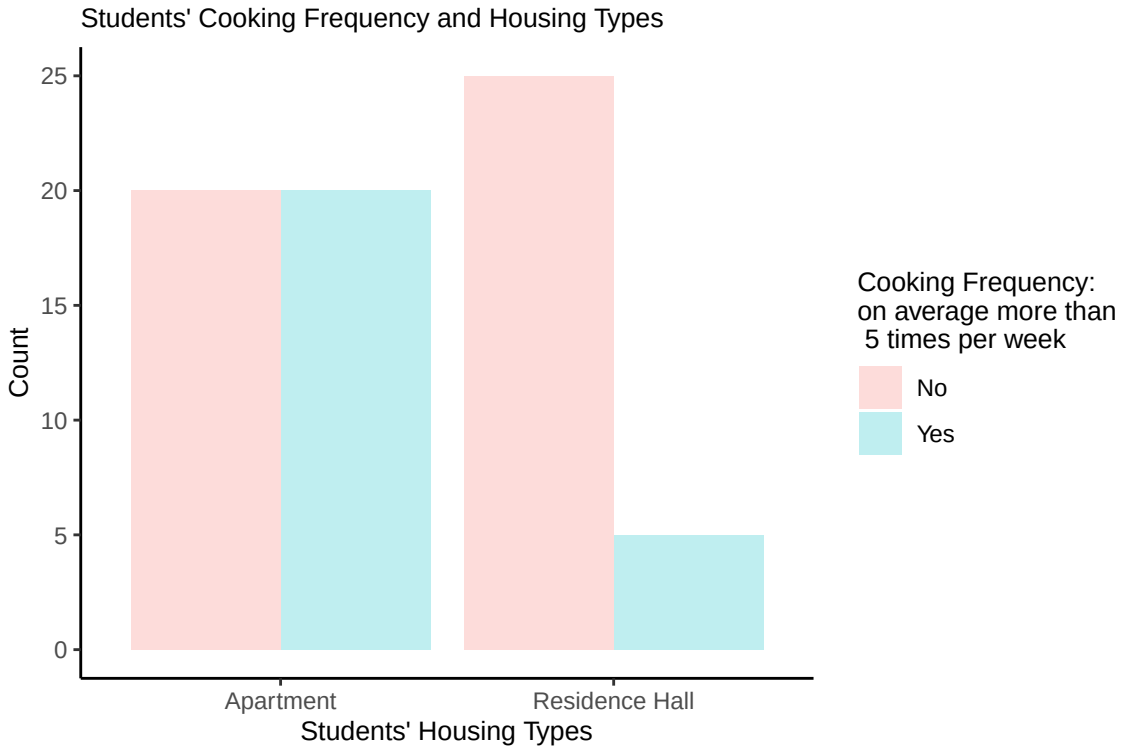
#### Contingency Table

| | Often Cook | Not Often Cook | Total |
|---|---|---|---|
| Students live in residence halls | 5 | 25 | 30 |
| Students live in apartments | 20 | 20 | 40 |
| Total | 25 | 45 | 70 |

#### Hypothesis

$$H_0 : Students' \ cooking \ frequency \ is \ not \ associated \ with \ their \ housing \ type$$
$$H_a : Students' \ cooking \ frequency \ is \ associated \ with \ their \ housing \ type$$

**Summary Figure**




**Checking for Assumption**

Using the equation of

$$Expected\ count\ = \frac{Row\ total * Column\ total}{Sample\ size}$$

| | Observed Count | Expected Count |
|---|---|---|
| Students live in residence halls and often cook | 5 | 10.71 |
| Students live in residence halls and do not often cook | 25 | 19.29 |
| Students live in apartments and often cook | 20 | 14.29 |
| Students live in apartments and do not often cook | 20 | 25.71 |

Since every expected count is larger than 5, so $\chi^2$ distribution could be used to test the hypothesis.

**Test Statistics**

$$\begin{aligned} X^2 &= \sum_{i=1}^{k} \frac{(x_i - np_{0i})^2}{np_{0i}} \\ &= \frac{(5-10.71)^2}{10.71} + \frac{(25-19.29)^2}{19.29} + \frac{(20-14.29)^2}{14.29} + \frac{(20-25.71)^2}{25.71} \\ &= 3.04 + 1.69 + 2.28 + 1.27 \\ &= 8.28 \end{aligned}$$

**Calculating p-value**

$$\begin{aligned} X^2 &\sim \chi^2_{(r-1)(c-1)} \\ p - value &= P(\chi^2_{(r-1)(c-1)} \geq X^2) \\ &= P(\chi^2_1 \geq 8.28) \\ &= 0.004 \end{aligned}$$

**Interpretation**

There is significant evidence that students' cooking frequency is associated with their housing choice (chi-square test for association, $\chi^2$=8.28, df=1, p=0.004, $\alpha$=0.05).

## 4.4 Linear Regression: Number of Times Eating in the Dining Halls and Cost of Housing

The variables of this test are Dining, which is defined as the average number of times a student eats in the dining halls per week, and Cost, which is the average cost of housing. The question derived from the research question of this project is whether there is a correlation between the cost of housing and the number of times eating in the dining hall per week. The hypothesis in this project is that there should be a positive relationship between those two variables, since students eating in the dining halls are more likely to be students living in the residence halls, which cost more than living in the apartments. Hence the aim of this test is to test whether there is a correlation between the number of times eating in the dining hall per week and the cost of housing.

**Hypotheses:**

The model used here to test the relationship between the average times of the student eating in the dining halls and the cost of housing is a linear model defined as follows:

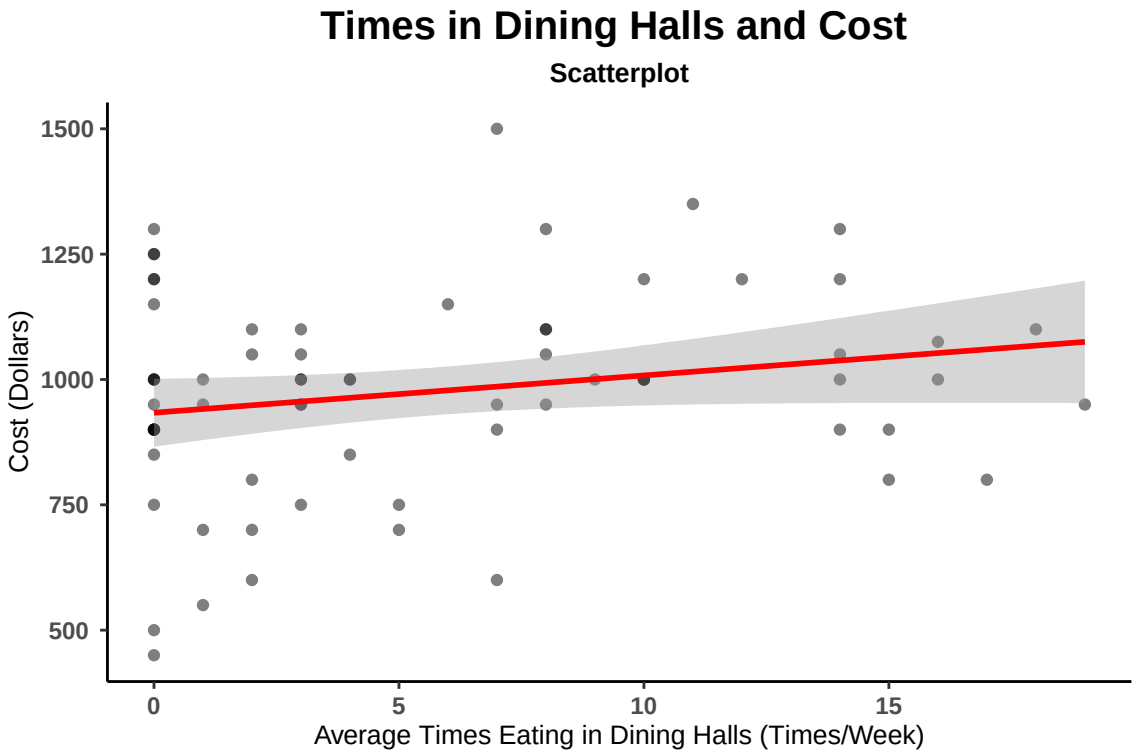
$$Y_i = \beta_0 + \beta_1 X_1 + \varepsilon$$

Where X is the explanatory variable number of times eating in the dining hall per week, Y is the response variable cost of housing, and $\varepsilon \sim N(0, \sigma^2)$ for some standard deviations. Test if the slope between the number of times eating in residence halls and the cost of housing is differenct from zero.

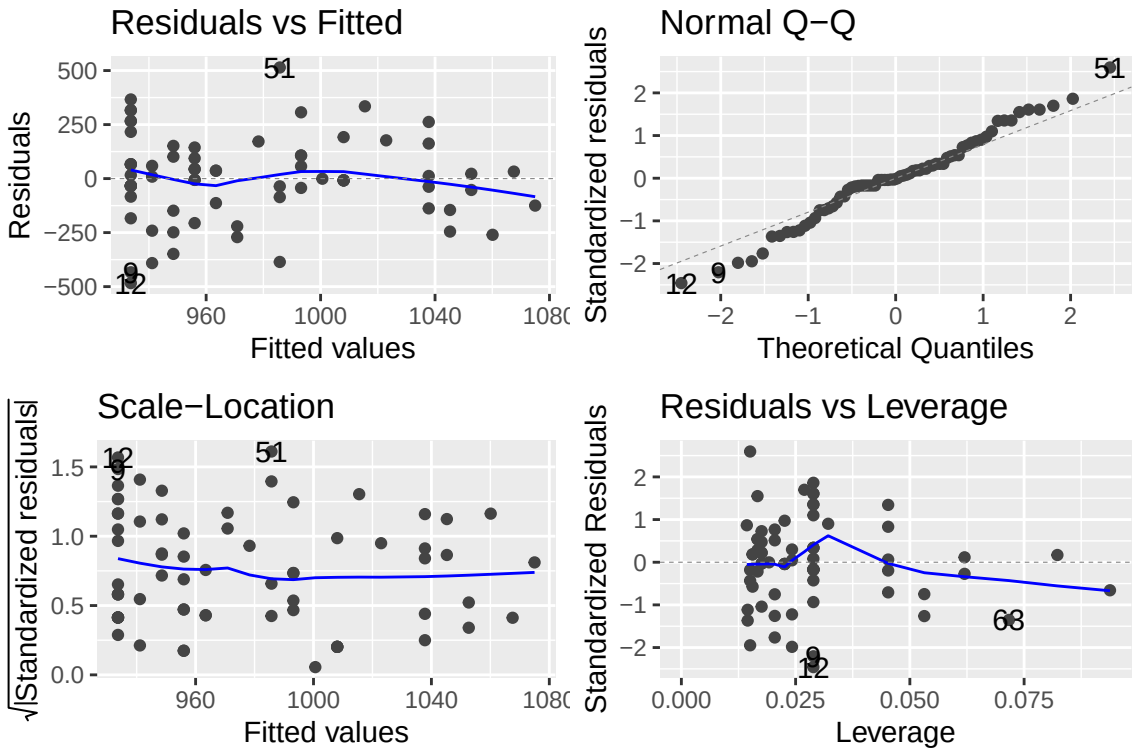
$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_a &: \beta_1 \neq 0 \end{aligned}$$

Where $\beta_1$ is the slope of the least squares line to predict the cost of housing based on the times eating in the residence hall per week.

**Summary Figure**




**Checking Assumptions**



The linearity assumption can be checked by the Residuals vs Fitted plot (top-left). There is no obvious fitted pattern via residual plot. The trend line is approximately horizontal at y = 0. Therefore, the linear assumption is satisfied. The normality assumption can be checked using the Normal Q-Q plot (top-right). The normal probability plot of residuals approximately follows the line y = x, although there are some points not very close to the line, but overall the trend is around y = x. Therefore, the normality assumption is satisfied The constant variance assumption can be checked by the Scale-Location plot (bottom-left). There is no obvious fitted pattern via scale-location plot. However, the trend line is decreasing slightly with respect to the fitted values over the range 900 to 1000. Therefore, the constant variance assumption may be violated. After inspecting the original scatterplot, we proceed and note the source of error. The sample size is not too small, and both the normal assumption and the equal variance assumption are met, hence a F-dsitribution is appropriate to use for testing here.

**Calculate Test Statistic**

Table 5: Table of Coefficients

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 933.674364 | 33.931298 | 27.516612 | 0.0000000 |
| dining | 7.438082 | 4.232447 | 1.757395 | 0.0833513 |

t* computed by using t-test for the slope between the number of times eating in the dining hall per week and the cost of housing.

$$t = \frac{b_1 - 0}{SE} = \frac{b_1}{SE} = \frac{7.438082}{4.232447} = 1.757395$$

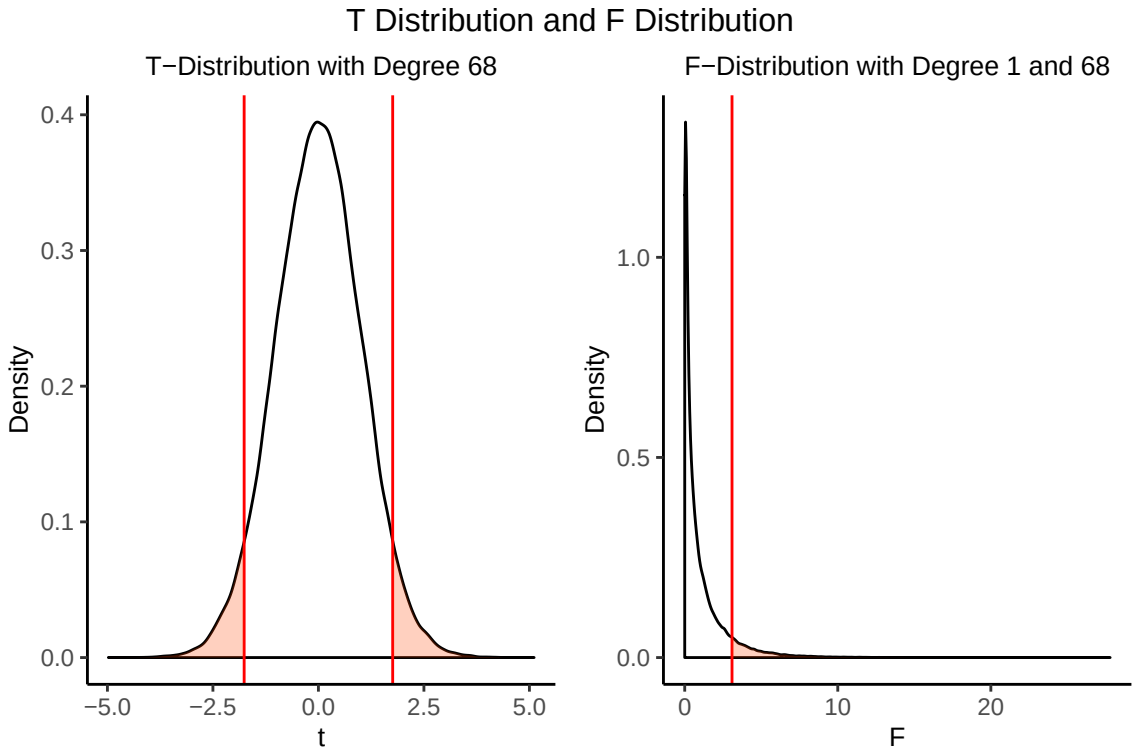
The ANOVA table is shown below.

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| dining | 1 | 123081.7 | 123081.7 | 3.088438 | 0.0833513 |
| Residuals | 68 | 2709963.0 | 39852.4 | NA | NA |

F-statistic is also computed from Anova table.

$$F = \frac{MSModel}{MSE} = \frac{123081.7}{39852.4} = 3.088438$$

**Compute P-Value**




For testing t,

$$T \sim t_{n-2}$$
$$p - value = P(|T| \geq |t|) = P(|T| \geq 1.76) = 0.083$$

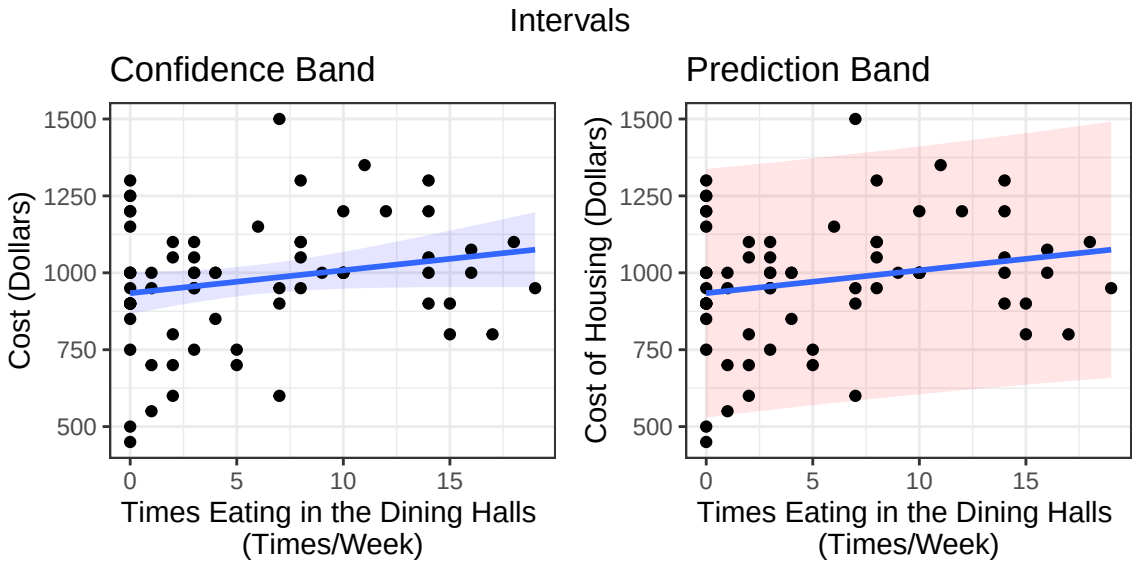
The p-value is calculated by finding the area under the t-distribution with the degrees of freedom (n-2)= 68, to the left of t=-1.76, and to the right of t=1.76.

For testing F,

$$F \sim F_{1,n-2}$$
$$p - value = P(F \geq 3.088438) = 0.083$$

The p-value is also calculated by finding the area under the F-distribution with the degrees of freedom 1 and 68, to the right of F=3.09. The p-values calculated by two tests are equavilent as expected.

#### Intervals




There is no significant evidence that there is a correlation between the number of times eating in the dining hall per week and the cost of housing. (two-tail t-test, df = 68, p = 0.083, $\alpha = 0.05$).

# 5. Discussion

## 5.1 Objectives

The objective of the study is to examine what factors influence the housing choices and campus life of Chinese students in UW-Madison. Therefore, multiple tests are utilized in this study to determine whether there is significant association between the aspects of students' college life, including employment, academics, demographics, and Chinese students' housing choices, either to live in a residence hall or apartment. The results of the tests reveal that having an on campus job or not doesn't affect the frequency of eating in the dining hall, and the cost of housing has significant relationship with housing choice.

## 5.2 Summary

The randomization test generally examines whether there is a difference between the number of times which students who do or do not have on campus jobs, eat in university dining halls. The variable of average number of times eating in dining halls in a week is a numeric variable, which indicates the popularity of dining halls in students' school life. It might consist of factors like the taste of the food or the convenience of eating in nearby halls. More importantly, it is reasonable to assume that students who work on campus may enjoy more favorable discounts on meals or have other benefits which lead them to choose to eat in the dining halls more often. The results of the test reject the hypothesis based on test statistics (sample difference of 0.06481481) and p-value of 0.9663. Since the p-value is large enough to exceed the significance level of 5%, the evidence of the existence of difference between the two groups is weak. Therefore, the conclusion of the randomization test is that regardless of whether students have campus jobs or not, they are likely to eat in the dining halls with similar weekly frequency. One possible implication is that the food served in the dining halls is popular among students, so whether or not they have discounts or other benefits of having a campus job does not influence their decision to eat there. However, another implication might be that students who work in the university do not receive favorable enough discounts in meal plans, compared to students who do not work on campus, that make them tend to eat more times in school's markets. If the latter situation is the case, the university may come up with more attractive policies which both encourage students to work in school and eat on campus.

The t test investigates the relationship between the cost variable and the housing variable. The cost is a quantitative variable representing how much money one individual would spend in rent and all utilities per month. For residence halls at UW-Madison, generally the cost is 800 to 1300 dollars per month; while for apartments, cost varies based on location and number of roommates and some other factors, generally it is between 500 and 1200 dollars per month. It is likely that cost is one factor that influences whether UW-Madison Chinese students choose to live in apartments or residence halls. Thus the null hypothesis is that the cost of living in an apartment is the same as the cost of living in a residence hall generally. From the t test, the p-value (0.008141) is less than the significance level of 0.05, so the test result is statistically significant. Therefore, there is strong evidence that lower cost of living in an apartment may be one factor that attracts UW-Madison Chinese students to live in apartments rather than residence halls.

The chi-square test investigates the association between two categorical variables: housing choice and cooking frequency. Housing choice is defined as where UW-Madison students live, either in residence halls or apartments. Living in apartments provides students more convenient access to kitchen, thus making cooking by themselves easier, while it is more difficult to cook in dorms. As a result, it is likely that housing choice is one factor that influences the cooking frequency of UW-Madison Chinese students choose to live in apartments and residence halls. Thus the null hypothesis is that there is no association between housing choice and cooking frequency of Chinese students in UW-Madison. From the chi-square test, the p-value 0.04 is less than the significance level of 0.05, so the test result is statistically significant. Therefore, there is strong evidence that preference of cooking by themselves may be one factor that attracts UW-Madison Chinese students to live in apartments rather than residence halls.

The linear regression examines the correlation between the times eating in the dining halls and the cost of housing. Based on the previous test, the average cost of housing of living in a residence hall is higher than

living in an apartment. Students living in residence halls are expected to eat more frequently in the dining halls, hence students who eat more frequently eating in the dining halls are expected to have higher cost of housing. However, from the t test and the F test, the p-value is 0.083 which is greater than the significance level of 0.05. Therefore, there is no significant statistical evidence that there is a nonzero correlation between these two variables.

In general, the analysis of the research shows that there are some connections between Chinese students' residential choice and campus life. Specifically, students who live in apartments often cook by themselves, and the mean cost of apartments is lower than the mean cost of dorms. Therefore, cooking environment and housing costs may be two of the factors which influences Chinese students' desicion about choosing where to live. Thus, to attract more Chinese students, residential halls may have to lower their prices or provide more spaces for cooking. However, there are also some disconnections shown from the results. Campus job has no influences on the number of times students eating in dining halls, which may infer that the dining benefits school provided for students employees are not favorable enough for them to eat more often in dining halls. Also, since the costs of housing is related with housing choices, the non-linear relationship between times eating in halls and the costs of housing may reveal that the students who live in apartments do not eat less often in dining halls than those who live in dorms. This could be interpreted as good news for the dining services: they may already offer tasteful enough food that attract students from all campus, or their convenient locations allow students to dine in more often. All in all, the results of the study could provide helpful guides for both future Chinese students and school's service institutions.

## 5.2 Error Analysis

One possible error is that the dataset is not selected randomly but using surveys on social media to collect volunteer responses, thus causing nonresponse bias as some students, who are in the friend circle of the researchers, may tend to fill out the survey while others, who are not friends with researchers, may not be represented in the results. Therefore, Chinese students who did not participate in the survey might be underrepresented. For instance, the friend circle of the researchers are mostly freshmen and sophomores. For those freshmen, they may choose to live in residence halls without taking other factors into consideration. Another possible error is response bias that students may tend to answer 0 for party frequency even if they go to parties to make them appear to be better students. In addition, stratified samples based on gender are used in the analysis but as the number of males who answered the survey is far lower than the number of females, a great amount of data gathered from female participants was missed after stratification, which causes the sample size to be relatively small so that the result is not representative.

## 5.3 Further Studies

Firstly, based on the result of the randomization test, having on campus job or not has no impact on students' eating times in dining halls, thus further studies could investigate what factors influencing students' decisions in their daily dining choice.

Secondly, based on the way this study was conducted, one suggestion for future study is to use random sampling methods, such as offline surveys, to ensure responses to be less biased and to produce more representative results and give more meaningful conclusions. Also, due to the limitation of data obtained for each variable, the limited time, and the limited knowledge, only four test methods learned from class were conducted. To fulfill the prerequisite of each test, only four hypotheses out of eight hypotheses are being tested. For instance, the original hypothesis for line regression is the number of credits a student takes is associated with the number of times a student eats in the dining hall. However, the data fail to achieve the normal assumption, so the alternative hypothesis is used, which might be less practically significant. With more time and resources, all of the eight hypotheses can be tested using more divergent test methods which may be fit better to the assumptions. Some hypotheses for future studies to test are: relationship between frequency of attending a party and the number of credits taken by the student, the

relationship between the housing choice and the distance from the student's living place to the lecture hall of his/her major, and the relationship between a student's housing choice and the number of credits he/she is taking. After performing all those tests, better suggestions could be given out to incoming freshmen Chinese Students to guide their housing choice, and school policy makers help improving the residence halls.

# 6. References

dplyr_0.8.3
ggplot2_3.2.1
tidyverse_1.2.1
grid_3.6.1
gridExtra_2.3
knitr_1.24
reshape2_1.4.3
ggfortify_0.4.8
xtable_1.8-4
R version 3.6.1 (2019-12-6)
RStudio (version 1.2.1335)

[1] Oyetunji, A ., & Abidoye, R. (2016). Assessment of the factors influencing students' choice of residence in Nigerian tertiary institutions. Sains Humanika 8(10). Retrieved from https://www.researchgate.net/publication/303739860_Assessment_Of_The_Factors_Influencing_Students'_Choice_Of_Residence_In_Nigerian_Tertiary_Institutions

[2] Kobue, T., Oke, A., & Aigbavboa, C. (2017). Understanding the determinants of students' choice of occupancy for creative construction. Procedia Engineering 196. Retrieved from https://www.sciencedirect.com/science/article/pii/S1877705817330916

[3] UT Synergy Journal. (n.d.). Away from Home: Chinese International Students at UW-Madison. Retrieved November 1, 2019,from http://utsynergyjournal.org/2019/03/25/away -from-home- chinese -international- students-at-uw-madison/

[4] University Housing. Retrieved from www.housing.wisc.edu

# 7. Appendix

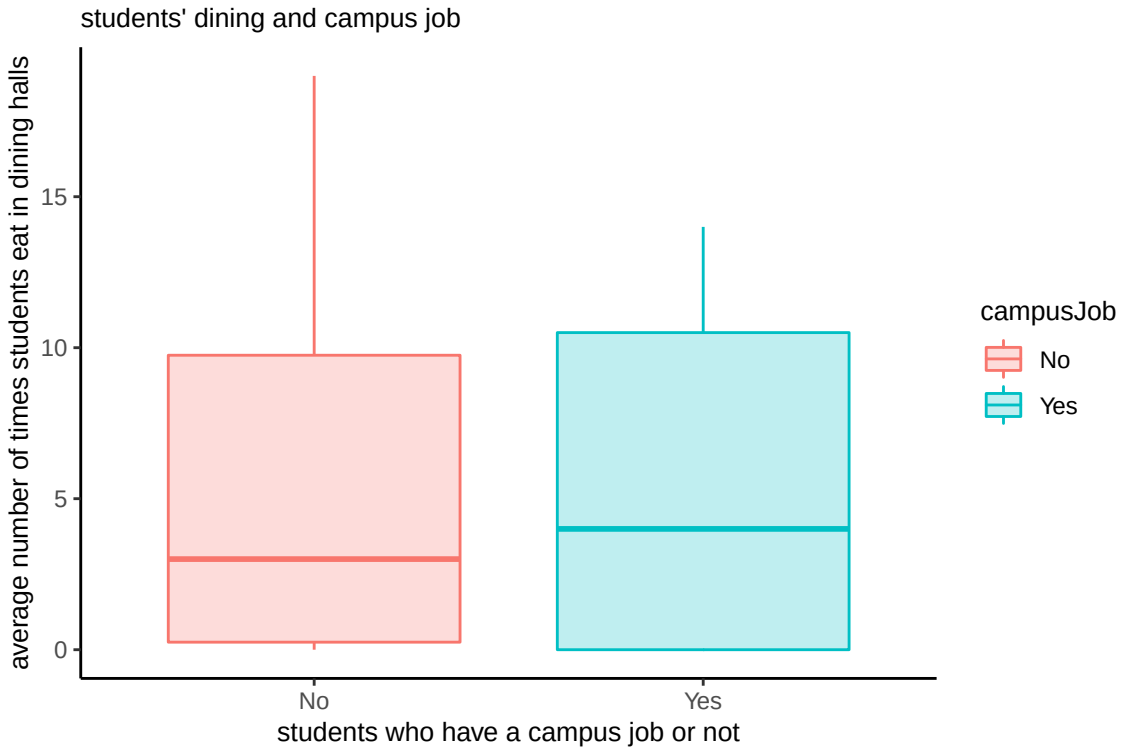
## 7.1 Packages

```
library(dplyr)
library(ggplot2)
library(tidyverse)
library(knitr)
library(grid)
library(gridExtra)
library(reshape2)
library(ggfortify)
library(xtable)
```

## 7.2 Randomization Test

### Summary Figure

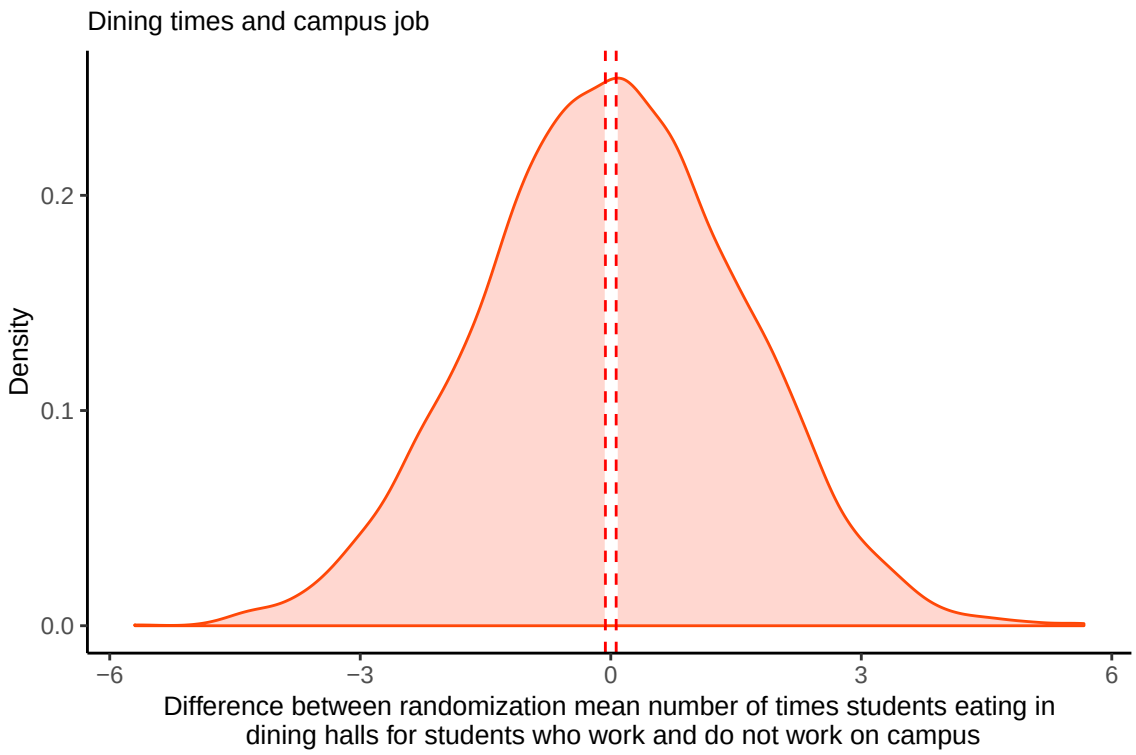
```
ggplot(data, aes(x = campusJob,y = dining)) +
geom_boxplot(aes(color=campusJob, fill=campusJob), alpha = 0.25) +
labs(x = "students who have a campus job or not",
y = "average number of times students eat in dining halls",
title = "students' dining and campus job") +
theme_classic() +
theme(plot.title=element_text(size=10),
axis.title=element_text(size=10),
legend.title=element_text(size=10))
```



**Checking Assumption**

```
dining.y <- data.hasJob %>%
  select(dining) %>%
  pull()
dining.n <- data.noJob %>%
  select(dining) %>%
  pull()

dining.y.bar <- mean(dining.y)
dining.n.bar <- mean(dining.n)
dining.bar <- (dining.y.bar + dining.n.bar)/2

shift.y <- dining.y.bar - dining.bar
shift.n <- dining.n.bar - dining.bar

dining.y.0 <- dining.y - shift.y
dining.n.0 <- dining.n - shift.n

set.seed(302)

B <- 10000
n.y <- length(dining.y)
n.n <- length(dining.n)

mat.rand.y <- matrix(sample(dining.y.0,B*n.y,replace=TRUE),
                     byrow = TRUE,
                     nrow = B,
                     ncol = n.y)
mat.rand.n <- matrix(sample(dining.n.0,B*n.n,replace=TRUE),
                     byrow = TRUE,
                     nrow = B,
                     ncol = n.n)
rand.mean.y <- apply(mat.rand.y,1,mean)
rand.mean.n <- apply(mat.rand.n,1,mean)
rand.diff <- rand.mean.y - rand.mean.n

df.rand <- data.frame(rand.diff)

gg<-ggplot(df.rand, aes(x=rand.diff)) +
  geom_density(color="orangered",alpha = 0.25) +
  geom_vline(xintercept=dining.y.bar-dining.n.bar, color="red",linetype="dashed") +
  geom_vline(xintercept=-(dining.y.bar-dining.n.bar), color="red",linetype="dashed") +
  labs(x = "Difference between randomization mean number of times students eating in\n dining halls for
       y = "Density",
       title = "Dining times and campus job") +
  theme_classic() +
  theme(plot.title=element_text(size=10),
        axis.title=element_text(size=10),
        legend.title=element_text(size=10))

d <- ggplot_build(gg)$data[[1]]

gg <- gg + geom_area(data = subset(d,x>dining.y.bar-dining.n.bar),aes(x=x,y=y),fill="tomato",alpha=0.25
```

```
  geom_area(data = subset(d,x<dining.n.bar-dining.y.bar),aes(x=x,y=y),fill="tomato",alpha=0.25)
gg
```




The randomization distribution is approximately symmetric and bell-shaped, so randomization test could be used.

**Test Statistics**

```
dining.y.bar - dining.n.bar
```

```
[1] 0.06481481
```

```
(ci.per.95 <- quantile(rand.diff, c(0.025,0.975)))
```

```
     2.5%     97.5%
-3.041667  3.023380
```

**Compute p-value**

$$P - Value = P(rand.diff \geq |\mu_1 - \mu_2|) = P(T \geq 0.06481481) = 0.9663$$

```
tol <- 1.0e-12
(p.value <- mean(abs(rand.diff) >= abs(dining.y.bar-dining.n.bar)-tol))
```

```
[1] 0.9663
```

### 7.3 T Test

```
data.2 <- read.csv("Data.csv")
colnames(data.2) <- c("Time", "Gender", "Age", "Grade", "Housing", "School",
                      "Job", "OnCamp", "Bulding", "Distance", "Cooking",
                      "Roommate", "Cost", "Credits", "Dining", "Party","N/A")
data.2 <- data.frame(data.2)

apt <- filter(data.2, Housing=="Apartment")
res <- filter(data.2, Housing=="Residence Hall")

cost.a <- apt$Cost
cost.r <- res$Cost

#Check for Assumptions
t <- data.2%>% group_by(Housing)%>% summarise(Count =n())
kable(t)
```

| Housing | Count |
|---|---|
| Apartment | 40 |
| Residence Hall | 30 |

```
#Summary Plot
##Boxplot
p1 <- ggplot(data.2, aes(x=Housing, y=Cost)) +
geom_boxplot(aes(color=Housing, fill=Housing, outlier.color=Housing), alpha = 0.25,
             outlier.alpha = 0.5, outlier.shape = 19, outlier.size = 1.5, outlier.stroke = 0.25) +
theme_classic() +
labs(x = "Housing Choice",
     y = "Cost ($)") +
theme(plot.title = element_text(size=10), axis.title = element_text(size=10), legend.title = element_te
theme(legend.position="none")

##Histogram
p2 <- ggplot(data.2, aes(x=Cost)) +
geom_histogram(aes(color=Housing, fill=Housing), bins = 30, boundary = 0, alpha = 0.1) +
facet_wrap(Housing~., nrow = 2) +
labs(x = "Cost ($)", y = "Count") +
theme_classic() +
theme(plot.title = element_text(size=10), axis.title = element_text(size=10), legend.title = element_te
theme(legend.position="none")

##Density Plot
p3 <- ggplot(data.2, aes(x=Cost)) +
geom_density(aes(color=Housing, fill=Housing), alpha = 0.1) +
facet_wrap(Housing~., nrow = 2) +
labs(x = "Cost ($)", y = "Density") +
theme_classic() +
theme(plot.title = element_text(size=10), axis.title = element_text(size=10), legend.title = element_te
theme(legend.position="none")
```

```
grid.arrange(p1, p2, p3, ncol=3, nrow=1, top=textGrob("Housing Choice v.s. Cost of Housing"))
```


```
#Calculate Test Statistic
n.a <- length(cost.a)
n.r <- length(cost.r)
x.a <- mean(cost.a)
x.r <- mean(cost.r)
s.a <- sd(cost.a)
s.r <- sd(cost.r)
t <- (x.a-x.r)/(sqrt(s.a^2/n.a+s.r^2/n.r))
t
```

```
[1] -2.550911
```

```
#Compute P-Value
p.value <- pt(t,min(n.a-1, n.r-1))
p.value
```

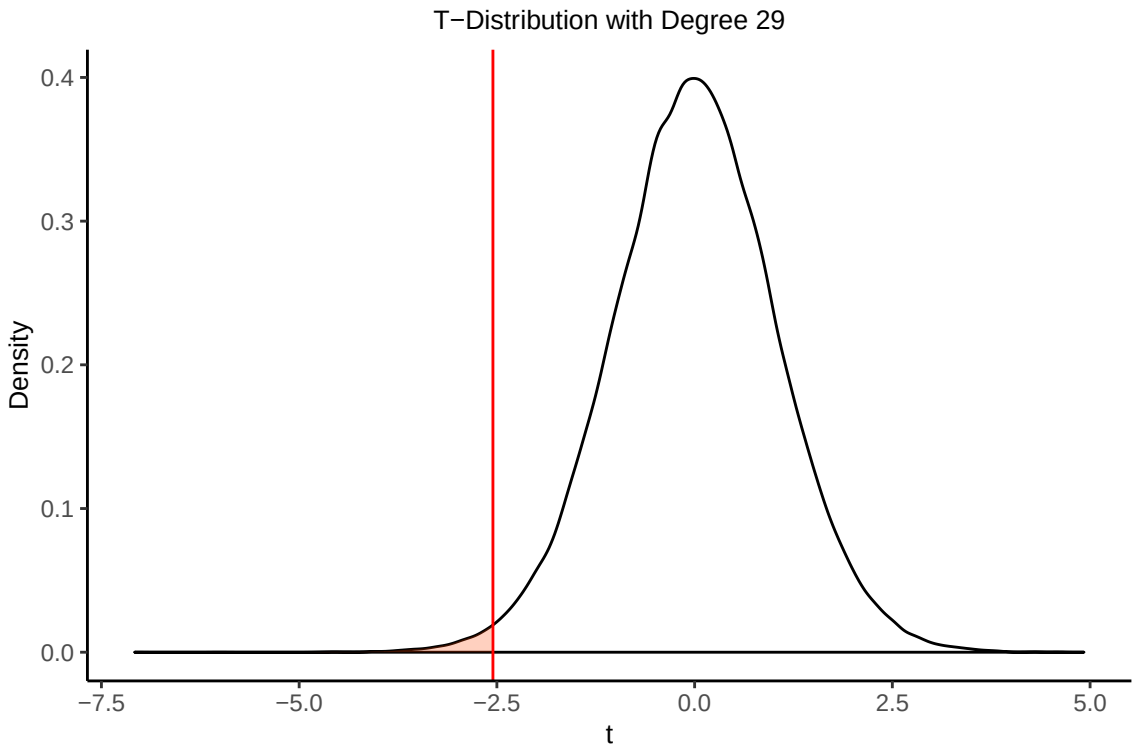
```
[1] 0.008141379
```

```
t.test(cost.a, cost.r, alternative = "less", conf.level = 0.95)
```

```
    Welch Two Sample t-test

data:  cost.a and cost.r
```

```
t = -2.5509, df = 66.093, p-value = 0.006534
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -39.21593
sample estimates:
mean of x mean of y
  927.500  1040.833
```

```
#Plot P-Value
t<-rt(100000, df=29)
t<-data.frame(t)
gg2 <- ggplot(t, aes(x=t)) + geom_density() +
  geom_vline(xintercept=-2.55, color="red") +
  theme_classic() +
  labs(x = "t", y = "Density", title = "T-Distribution with Degree 29")+
  theme(plot.title=element_text(hjust = 0.5)) +
  theme(plot.title = element_text(size=10), axis.title = element_text(size=10),
        legend.title = element_text(size=10))
d <- ggplot_build(gg2)$data[[1]]

gg2 <- gg2 + geom_area(data = subset(d,x<=-2.55),aes(x=x,y=y),fill="orangered",alpha=0.25)
gg2
```




```
#Compute 95% Confidence Interval
t.test(cost.a, cost.r, alternative = "two.sided", conf.level = 0.95)
```

```
    Welch Two Sample t-test
```

```
data:  cost.a and cost.r
t = -2.5509, df = 66.093, p-value = 0.01307
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -202.03552  -24.63114
sample estimates:
mean of x mean of y
  927.500  1040.833
```

## 7.4 Chi-square test

### Calculating Observed Count

```
data <- read.csv("Data.csv",header=T)
colnames(data) <- c("date","gender","age","grade","housing type","school","job","campusJob","building",

hall.often.cook <- data%>%
  filter(`cook`=="Yes"& `housing type`=="Residence Hall")
nrow(hall.often.cook)
```

```
[1] 5
```

```
hall.notoften.cook <- data%>%
  filter(`cook`=="No"& `housing type`=="Residence Hall")
nrow(hall.notoften.cook)
```

```
[1] 25
```

```
apt.often.cook <- data%>%
  filter(`cook`=="Yes"& `housing type`=="Apartment")
nrow(apt.often.cook)
```

```
[1] 20
```

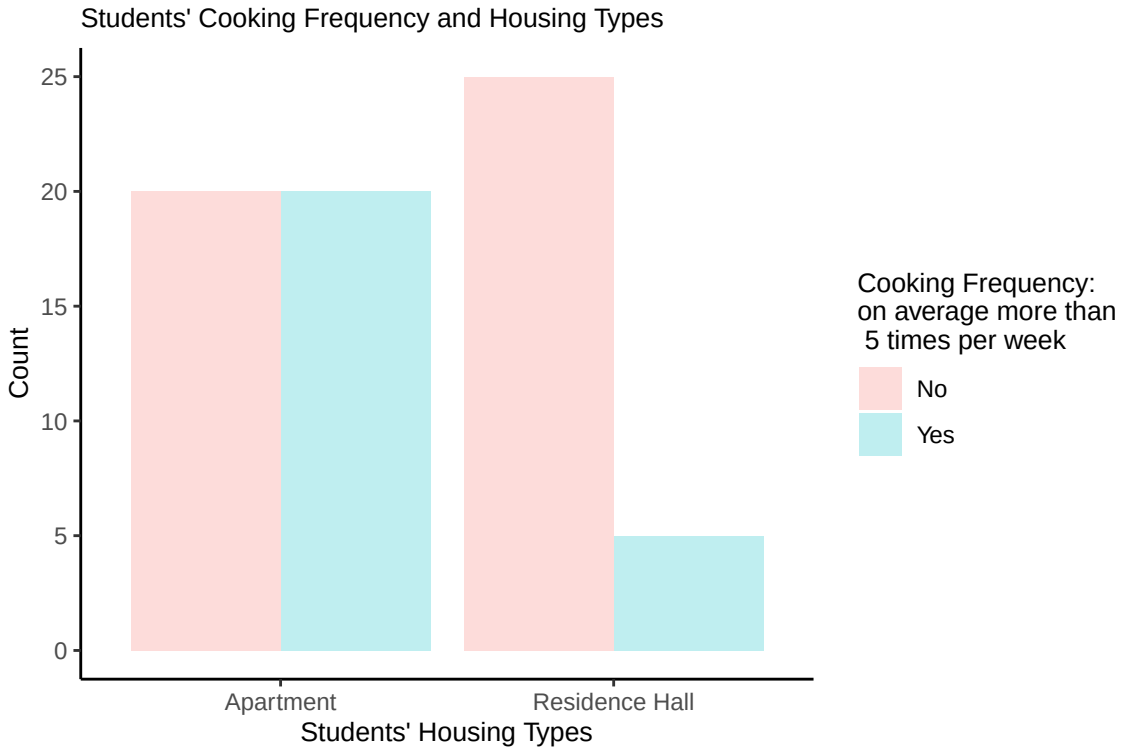
```
apt.notoften.cook <- data%>%
  filter(`cook`=="No"& `housing type`=="Apartment")
nrow(apt.notoften.cook)
```

```
[1] 20
```

### Summary Figure

```
t <- table(data$`housing type`,data$cook)

melt(t)%>%
  ggplot(aes(x=Var1,y=value,fill=Var2)) +
  geom_bar(position="dodge",stat="identity",alpha = 0.25) +
  labs(x = "Students' Housing Types",
```

```
       y = "Count",
       title = "Students' Cooking Frequency and Housing Types",
       fill = "Cooking Frequency: \non average more than\n 5 times per week",
       color = "Homicide Frequency Level") +
  theme_classic() +
  theme(plot.title=element_text(size=10),
  axis.title=element_text(size=10),
  legend.title=element_text(size=10))
```




### Calculating p-value

```
t <- table(data$`housing type`,data$cook)

pchisq(8.28,1,lower.tail=FALSE)
```

```
[1] 0.004008413
```

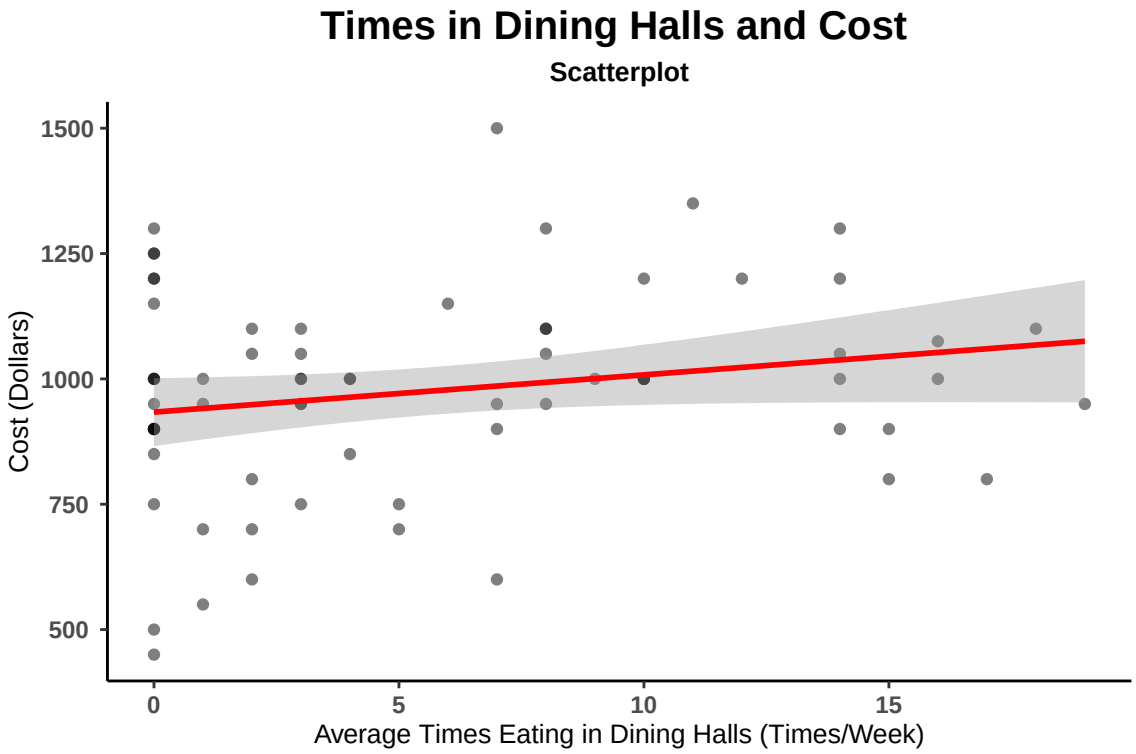
```
chisq.test(t, correct = FALSE)
```

```
    Pearson's Chi-squared test

data:  t
X-squared = 8.2963, df = 1, p-value = 0.003973
```

## 7.5 Linear Regression

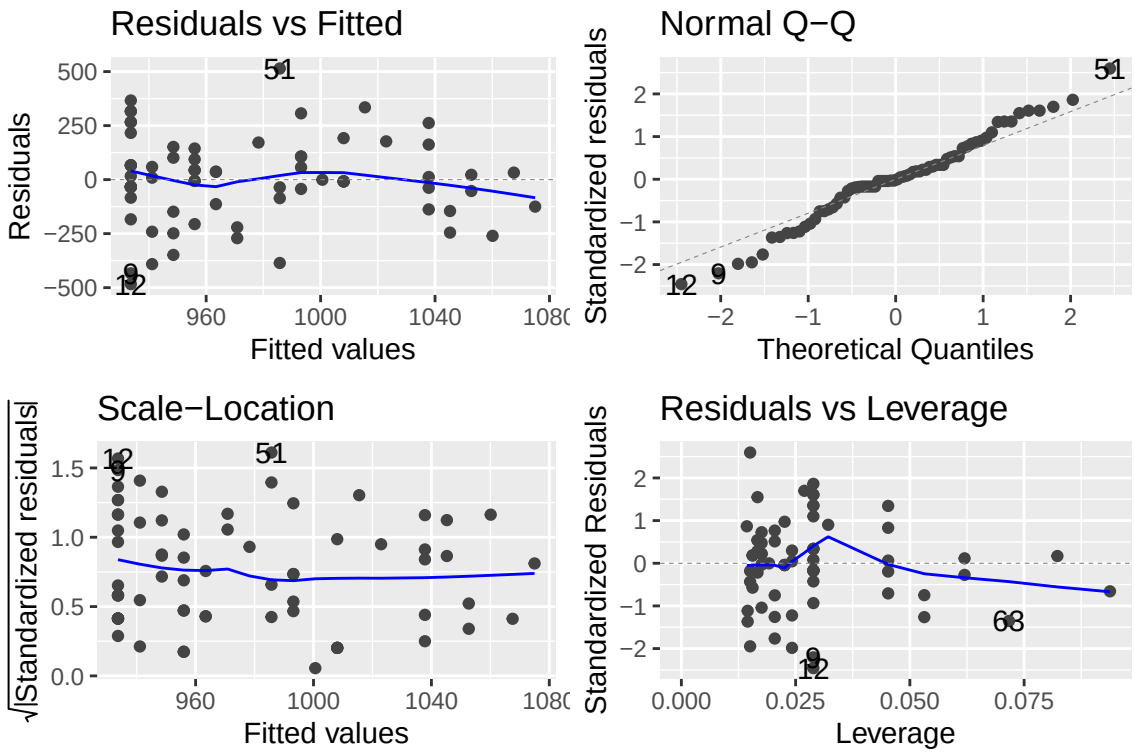
### Summary Figure

```
data.2 <- read.csv("Data.csv")
colnames(data.2) <- c("time", "gender", "age", "grade", "housing", "school",
                      "job", "oncamp", "building", "distance", "cooking",
                      "roommate", "cost", "credits", "dining", "party","X")
data.2 <- data.2 %>%
  select(-17)

ggplot(data.2, aes(dining, cost))+geom_point(alpha = 0.5)+
geom_smooth(method = "lm", color = "red") +theme_classic() +
labs(subtitle = "Scatterplot", title = "Times in Dining Halls and Cost ",
x = "Average Times Eating in Dining Halls (Times/Week)", y = "Cost (Dollars)") + theme(plot.title = ele
color = "black", hjust = 0.5, face = "bold"), plot.subtitle = element_text(size = 10,
color = "black", hjust = 0.5, face = "bold"), axis.title.x = element_text(size = 10),
axis.title.y = element_text(size = 10), axis.text.x = element_text(vjust = 0.6,
face = "bold"), axis.text.y = element_text(hjust = 0.6, face = "bold"))
```




### Checking Assumptions

```
lm.dining <- lm(cost ~ dining, data.2)
mod.dining <- summary(lm.dining)
autoplot(lm.dining)
```




**Calculate Test Statistic**

```
xtable(mod.dining) %>%kable(caption = "Table of Coefficients")
```

Table 8: Table of Coefficients

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 933.674364 | 33.931298 | 27.516612 | 0.0000000 |
| dining | 7.438082 | 4.232447 | 1.757395 | 0.0833513 |

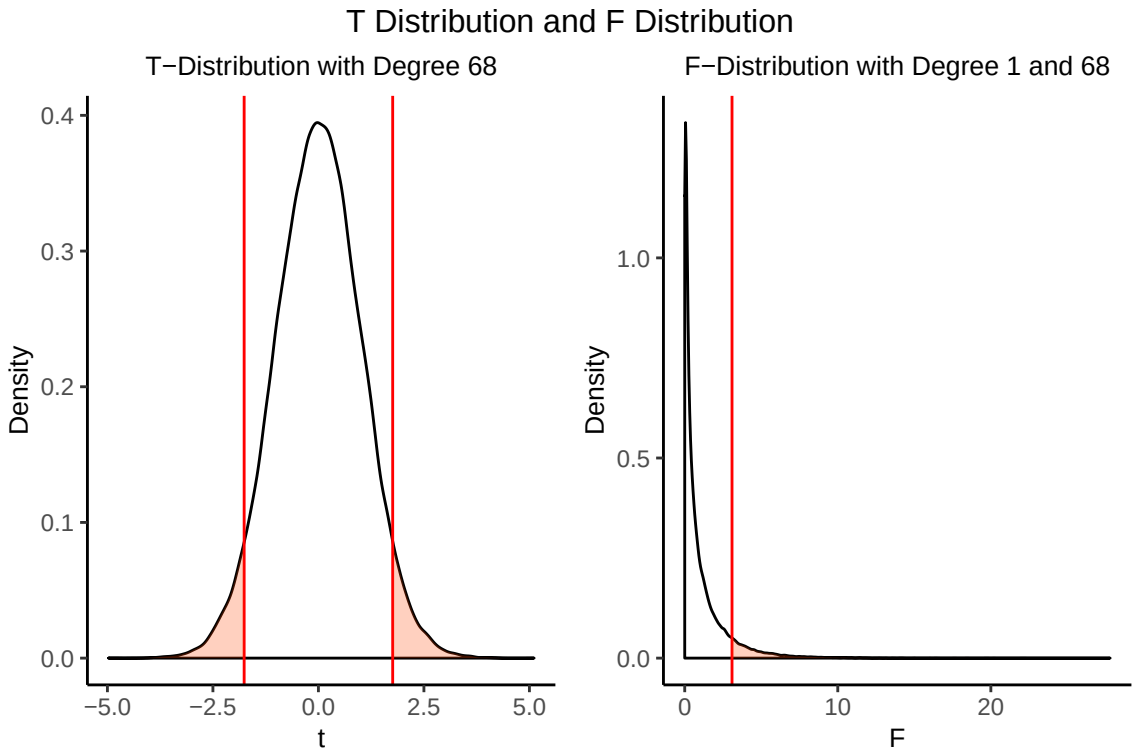
```
xtable(anova(lm.dining)) %>%
  kable()
```

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| dining | 1 | 123081.7 | 123081.7 | 3.088438 | 0.0833513 |
| Residuals | 68 | 2709963.0 | 39852.4 | NA | NA |

**Compute P-Value**

```
#Plot P-Value
t<-rt(100000, df=68)
t<-data.frame(t)
gg3 <- ggplot(t, aes(x=t)) + geom_density() +
  geom_vline(xintercept=c(-1.76,1.76), color="red") +
```

```
  theme_classic() +
  labs(x = "t", y = "Density", title = "T-Distribution with Degree 68")+
  theme(plot.title=element_text(hjust = 0.5)) +
  theme(plot.title = element_text(size=10), axis.title = element_text(size=10),
        legend.title = element_text(size=10))
d <- ggplot_build(gg3)$data[[1]]
gg3 <- gg3 + geom_area(data = subset(d,x<=-1.76),aes(x=x,y=y),fill="orangered",alpha=0.25)
p6 <- gg3 + geom_area(data = subset(d,x>=1.76),aes(x=x,y=y),fill="orangered",alpha=0.25)

#Plot F Statistic
f <- rf(100000, df1=1, df2=68)
f <- data.frame(f)
gg4 <- ggplot(f, aes(x=f)) + geom_density() +
  geom_vline(xintercept=3.09, color="red") +
  theme_classic() +
  labs(x = "F", y = "Density", title = "F-Distribution with Degree 1 and 68")+
  theme(plot.title=element_text(hjust = 0.5)) +
  theme(plot.title = element_text(size=10), axis.title = element_text(size=10),
        legend.title = element_text(size=10))
d <- ggplot_build(gg4)$data[[1]]
gg4 <- gg4 + geom_area(data = subset(d,x>=3.09),aes(x=x,y=y),fill="orangered",alpha=0.25)
p7 <- gg4
grid.arrange(p6, p7, ncol=2, nrow=1, top=textGrob("T Distribution and F Distribution"))
```




**Intervals**

```r
## Confidence Interval for $\hat{y}$
conf <- predict(lm.dining,data,interval="confidence")
conf <- data.frame(conf)
df <- data %>%
  mutate(lwr.ci = conf$lwr, upr.ci = conf$upr)
p4 <- ggplot(df, aes(x=dining,y=cost,ymin=lwr.ci,lwr,ymax=upr.ci)) +
    geom_ribbon(fill="blue",alpha=0.1) +
    geom_point() +
    geom_smooth(method="lm",se=FALSE) +
    xlab("Times Eating in the Dining Halls \n(Times/Week)") +
    ylab("Cost (Dollars)") +
    ggtitle("Confidence Band") +
    theme_bw()

## Prediction Interval for $\hat{y}$
pred <- predict(lm.dining,data,interval="predict")
pred <- data.frame(pred)
df <- data.2 %>%
  mutate(lwr.pred = pred$lwr, upr.pred = pred$upr)
p5 <- ggplot(df, aes(x=dining,y=cost,ymin=lwr.pred,ymax=upr.pred)) +
    geom_ribbon(fill="red",alpha=0.1) +
    geom_point() +
    geom_smooth(method="lm",se=FALSE) +
    xlab("Times Eating in the Dining Halls \n(Times/Week)") +
    ylab("Cost of Housing (Dollars)") +
    ggtitle("Prediction Band") +
    theme_bw()

grid.arrange(p4, p5, ncol=2, nrow=1, top = textGrob("Intervals"))
```