Middlebury Sophomore Housing Selection Process

The Trash CAN: Conor Donahue, Aidan Garcia, Naomi Atwood

2025-05-18

Introduction

Wrapping up our first year of college at Middlebury, preparations for next year have already begun: registering for classes, organizing storage units, and selecting housing for the 2025-2026 school year. Very different from our first year (random assignment), sophomore housing selection was a very stressful and overwhelming process for some, or full of joy and excitement for others. Students choose a housing group (of size 1 to 8 people) and each group is assigned a time slot for room selection based on a random lottery system. We wanted to explore the results of this housing process for the class of 2028/.5. We will do this by investigating three research questions:

1) What is the true proportion of Middlebury 2028/.5 students who wanted each housing type (suite, double, or single)?
2) Is satisfaction with the housing selection process higher among students who had an earlier time slot?
3) Is the proportion of people satisfied with their housing selection less for those who didn't receive their desired housing compared to those who did?

Data Collection

To collect our data, we distributed a Google Form across campus to try to reach the class of 2028/.5. The Google Form gathered information on 8 variables:

- Participation in the lottery housing selection process (if no, must remove from data set)
- Size of housing group (1-8)

- First choice living situation (suite, double, single)
- Actual living situation (suite, double, single)
- Building (Pearsons, Milliken, Hadley, Gifford, Coffrin)
- Satisfaction with housing situation (1-100)
- Satisfaction with selection process (1-100)
- Housing draw time slot (1-225)

This form was distributed in many different ways. We sent a link to the survey in the Middlebury Mountain Club and the Middlebury Pranksters GroupMe. We also handed out QR codes of the survey to people outside of Proctor Dining Hall one day at lunch, and to people at the Freshman Dinner in Atwater Dinning Hall. The survey was also posted in some of the freshman residence halls. These sampling methods were form a convenience sampling.

Methods

Question 1:

prop_double	prop_single	prop_suite
0.1044776	0.2835821	0.5970149

The table above shows the observed proportions of students who wanted each housing option: suite, double, and single. To determine the true proportion of Middlebury class of 2028/.5 students who wanted each housing option (suite, double, or single), given by the 'first choice living situation' variable, we calculated a 95% confidence interval. For each housing option, we simulated the housing sample data 5000 times using a bootstrap method. We labeled 67 cards with all of our observed outcomes and randomly drew 67 with replacement. We then calculated the proportion of students who wanted a suite from that bootstrap sample. We repeated this process 5000 times until we created a sampling distribution of the proportion of students who wanted a suite as their first choice. We repeated this bootstrapping process for the proportion of students who wanted a double and then a single. The bootstrapping method was effective because it showed us how the data behaved with many, many trials. Next, we calculated a 95% confidence interval for each sampling distribution. We did this by finding the upper 97.5% and lower 2.5% to determine the center 95% on each graph. These points gave us the lower and upper bounds for our confidence interval.

Question 2:

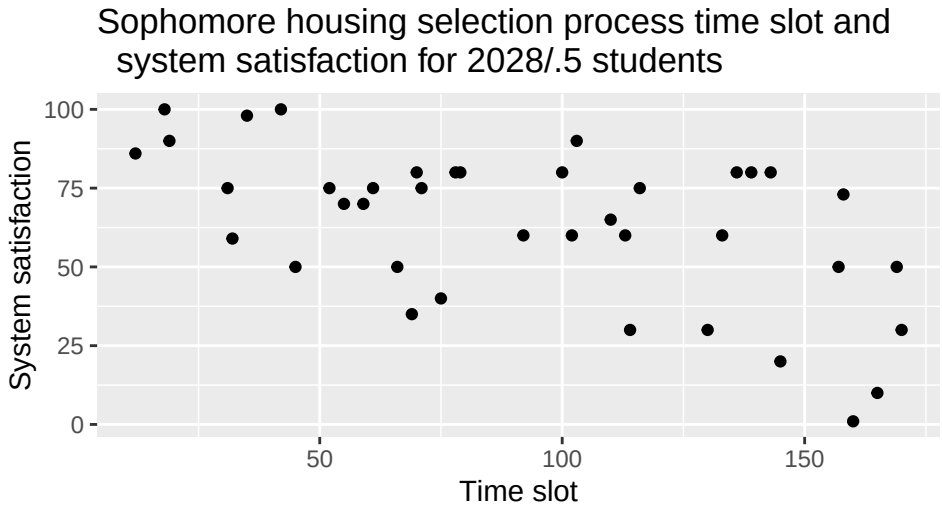
We were motivated to investigate question 2 because we hypothesized that students who randomly received an early time slot would be more satisfied with the housing selection process in general because they benefited from the system. To answer our question, we will focus on the time slot and system satisfaction variables. Because both of these variables are numerical, we decided to fit a linear regression to the data and perform a hypothesis test to see if the slope is significantly non-zero.

$system\ \widehat{satisfaction} = \beta_0 + \beta_1 * time\ slot + \epsilon$

Before we can start testing our data, we must wrangle the data so that it is independent. We received many responses that had the same time slot, meaning that individuals were in the same housing group. This violates independence because individuals with the same time slot are likely to have similar first choice and actual housing situations, housing group size, and satisfaction feelings about their situation and the system. We also had some non-response for the time slot because people couldn't remember or they didn't want to look it up. To remove any time slot duplicates and NAs, we filtered out the NAs in the time slot column and then grouped by time slot. Then we randomly sampled one of the rows from each time slot. This created a new data set with 39 independent observations.

The first step of a linear regression analysis is checking conditions for inference. We can start by making a scatter plot of the data with time slot on the x-axis and system satisfaction on the y-axis. This helps check our condition of linearity because it appears that the data has a rough negative linear relationship. We know that the data is independent because we just removed any time slot duplicates, and the our data was collected in a random manner. Because the first two conditions are met, we can continue with fitting a linear regression.

We can start by passing the independent data into the lm() function, fitting a linear regression to our data. Then we can pull the b_0 and b_1 coefficient estimates using the tidy() function and create our fitted model. Now we can check the rest of our conditions by plotting a histogram of the residuals to see if they are approximately normal and creating a scatter plot of the residuals to see if they have equal variance. If these conditions are met, we can continue with a hypothesis test at the $\alpha = 0.05$ significance level. We can let H_0: $\beta_1 = 0$: H_A be $\beta_1 < 0$. We can pull our test statistic and the p-value from the lm() function. The p-value in the table is automatically 2-sided so if our test statistic is in the direction of our hypothesis, then we can divide the p-value by 2 to find the p-value for our hypothesis. Based on the p-value, we can make a decision about the hypothesis and a conclusion about our research question.



Question 3:

outcome	house_satisfied	n	prob
No	Satisfied	9	0.7500000
No	Unsatisfied	3	0.2500000
Yes	Satisfied	32	0.9411765
Yes	Unsatisfied	2	0.0588235

To answer the research question, whether the proportion of Middlebury 2028/.5 students who were satisfied with their housing selection and did not get their desired housing option was less than that of those who did get their desired housing option, a difference in proportions hypothesis test was conducted. In this case, a "satisfied" person was considered one who responded with a housing satisfaction of 50 or less out of 100. Also, to ensure as much independence as possible, only one respondent from a given housing group (identified by the same time slot) was randomly selected, as people in the same housing groups are likely to have responses that would give information about others in the group. The Central Limit Theorem could not be used to obtain the null distribution because the success-failure condition was not met, so simulation methods were used instead. Essentially, 41 "cards" were labeled "Satisfied", while the remaining 5 were labeled "Unsatisfied." This is because in total, disregarding whether or not desired housing was received, 41 students were satisfied while only 5 were unsatisfied. These were then shuffled, and 12 were dealt out to the group who did not get desired housing (for n_n = 12), while 34 were dealt to the group who did (for $n_y = 34$). The proportion of "Satisfied" cards was then taken for each group, allowing for the difference in the two proportions to be taken. By shuffling the cards together, we centered the difference in proportions of students who were satisfied (did not receive - did receive desired housing) at 0 to simulate the null

hypothesis. This process was repeated B = 5000 times to obtain the null distribution. The p-value was calculated by finding the probability of getting the observed difference in proportion (-0.1911765), or less, given the null distribution is true.

p_y = Proportion of people who were satisfied with their housing selection, given that they got their first choice housing option.

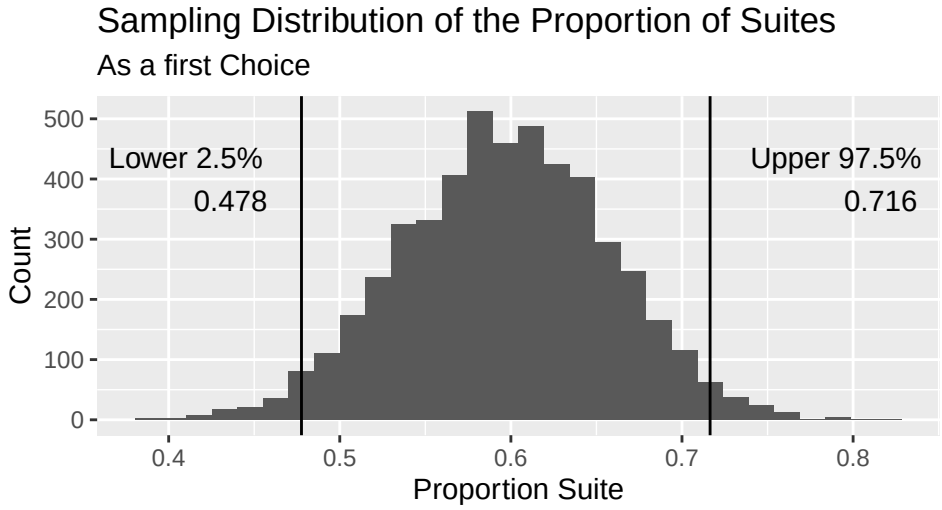
p_n = Proportion of people who were satisfied with their housing selection, given that they did not get their first choice housing selection.

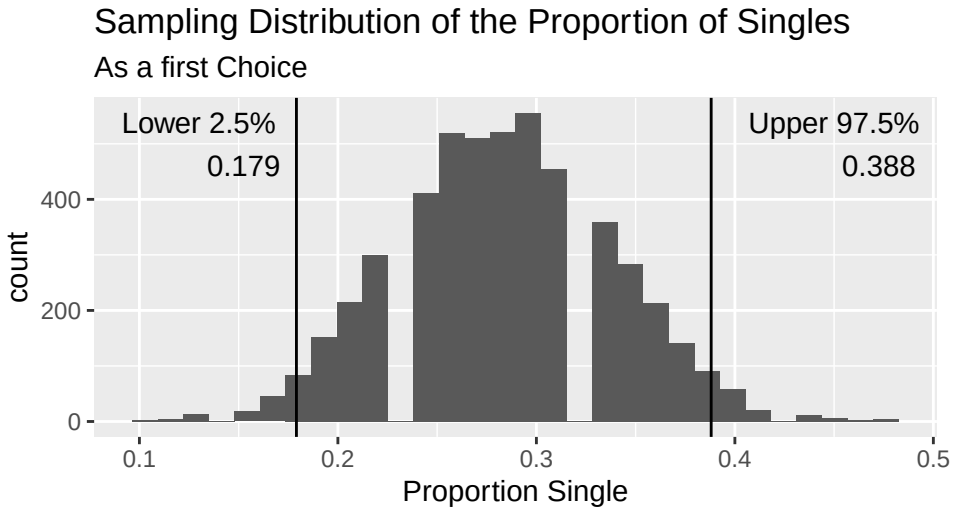
$H_0 : p_n - p_y = 0, H_A : p_n - p_y < 0$

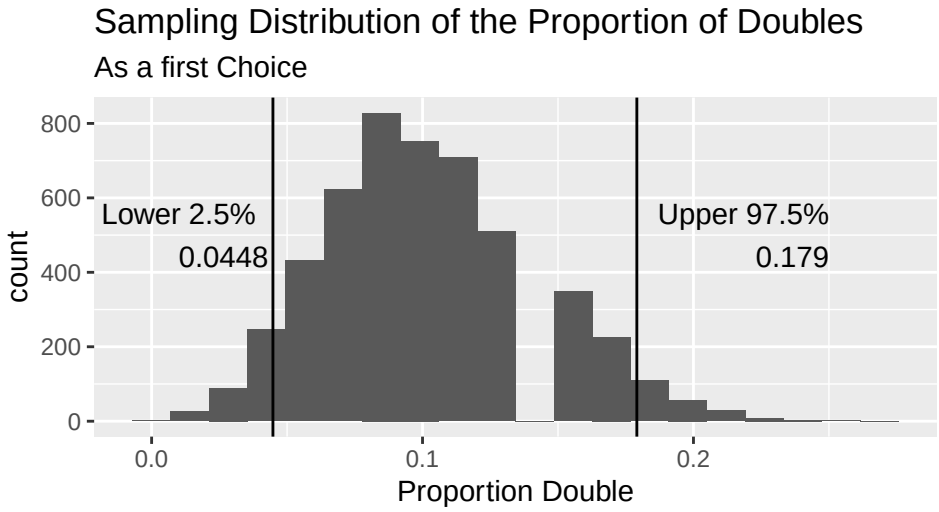
Significance Level: $\alpha = 0.05$

Results

Question 1:



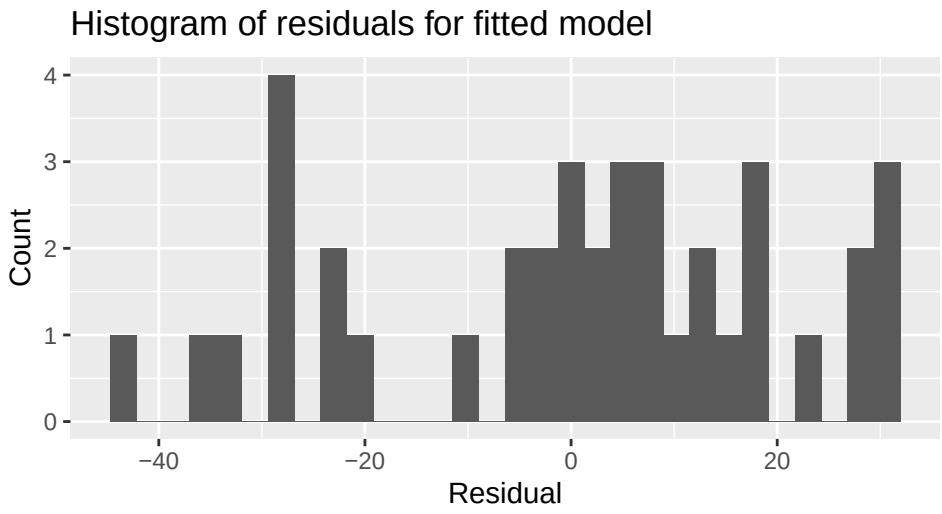




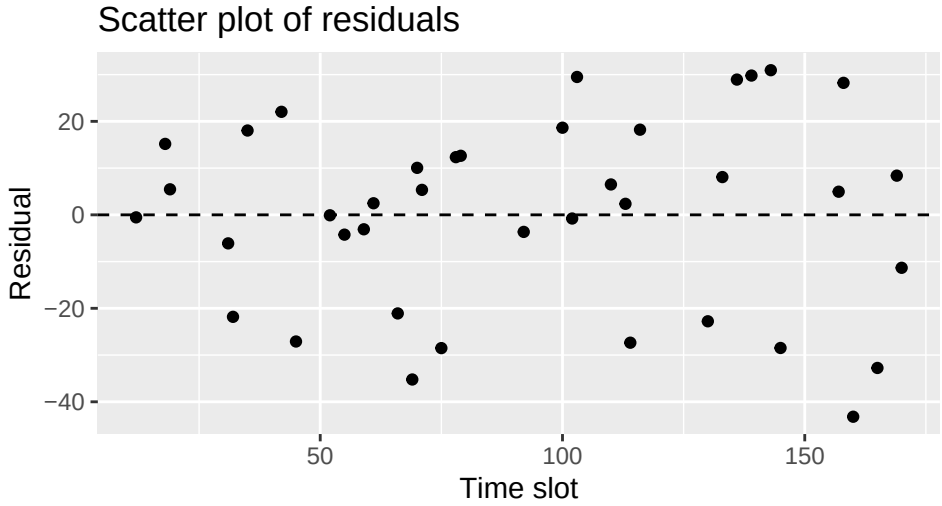
We are 95% confident that the true proportion of Middlebury class of 2028/.5 students who wanted a suite was between 0.4776119 and 0.7164179. We are 95% confident that the true proportion of Middlebury class of 2028/.5 students who wanted a single was between 0.1791045 and 0.3880597. Lastly, we are 95% confident that the true proportion of Middlebury class of 2028/.5 students who wanted a double was between 0.0447761 and 0.1791045.

Question 2:

The fitted model is $system\ \widehat{satisfaction} = 89.9746055 - 0.2861506 * time\ slot$.



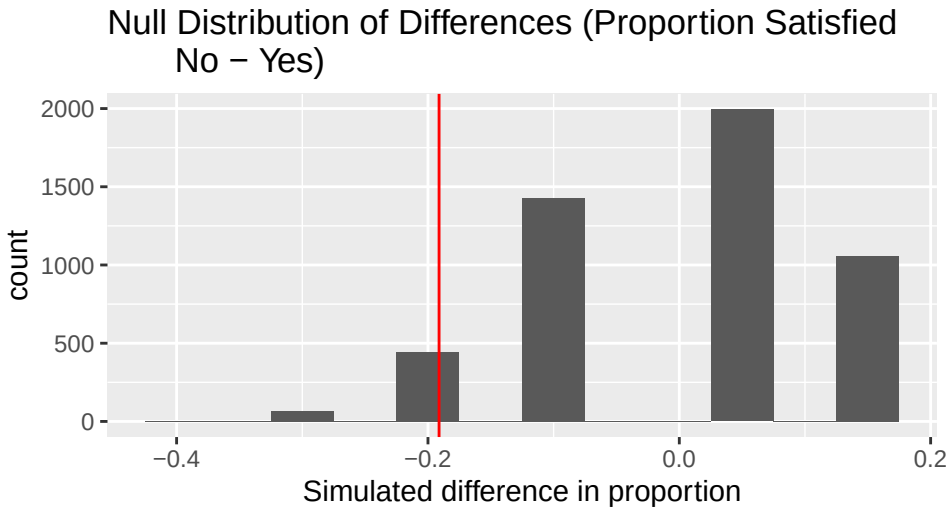
The histogram of the residuals for the fitted model is not exactly normal. It appears it could be skewed slightly left, but it is not very dramatic. We will proceed with the test, acknowledging that the residuals aren't exactly normal.



The scatter plot of residuals for the fitted line show approximately equal variance. There are no patterns that would indicate that a linear model is not the best model.

The test statistic for our fitted model is -4.065232, with a p-value of 1.2026343×10^{-4}. Because 1.2026343×10^{-4} is less than $\alpha = 0.05$, we reject H_0. There is sufficient evidence to suggest that β_1 is less than zero. This means that there is evidence of a negative linear relationship between time slot and housing selection system satisfaction because the slope of the SLR is less than 0.

Question 3:



Because our p-value of 0.1032 is greater than $\alpha = 0.05$ and we fail to reject H_0. This means there is not convincing evidence that the proportion of Middlebury 2028/.5 students who were satisfied with their housing selection, but did not get their desired housing, was less than the proportion of students who were satisfied given they received their desired housing selection.

Discussion

In summary, we estimated the true proportions of Middlebury College class of 2028/.5 students who wanted a single, double and suite for their housing for the 2025/2026 academic year using 95% confidence intervals. We are 95% confidant that the true proportion of students who wanted a suite was between 0.4776119 and 0.7164179, the proportion of students who wanted a single was between 0.1791045 and 0.3880597, and the proportion of students who wanted a double was between 0.0447761 and 0.1791045. This shows that almost a majority of students wanted a suite, consistent with our conversational observations. When discussing housing with our peers, we felt like most people we talked to expressed interest in wanting a suite as their ideal housing option.

We also found that there is evidence of a negative linear relationship between time slot and housing system satisfaction among 2028/.5 students. Using a hypothesis test for β_1, we showed that there is significant evidence that β_1 is nonzero, and thus there is a linear relationship between the variables. This aligns with our hypothesis because it shows that individuals who had an earlier time slot, and presumably got their desired rooming choice, were more satisfied with the selection system as a whole. This is because they benefited from the system, and thus approve of it.

Finally, we concluded that there is no difference in satisfaction between individuals who got their desired housing option and those who did not. Through a hypothesis test for a difference in proportions, we failed to reject H_0, meaning that there is not a statistically significant difference in satisfaction. This may suggest that overall satisfaction was relatively high, and that the housing selection system is pretty effective in that sophomores get housing they can be happy with. People are also living with or near their housing group regardless of what type of rooms they get, so satisfaction could be high across both these groups (did and didn't receive desired housing) as people are simply living with their friends.

When evaluating the validity of our data and analysis, it is important to consider how the data was obtained and how representative of our entire population it is. Our sampling method was mostly based on convenience. We distributed a Google form at some meals, in some residence halls, and sent it in some GroupMe chats. We didn't sample in a completely random manner which could have impacted our results. We also often distributed the form to a group of people that were already sitting together, or linked in some way. This means that many people from the same housing group filled out the survey. To account for this lack of independence, we had to remove people from the data set so that there were no time slot duplicates. This significantly shrunk our sample size, which could have also affected our results. For example, we didn't have very many observed individuals that didn't get their ideal housing choice so this could have skewed our results in question 3. We were also interested in researching if housing group size affected if individuals got their first choice housing option. Once we collected our data, we realized we couldn't answer this question because we didn't have enough data on people who didn't get their first choice option. A larger sample size and completely random sampling could help improve our results and analysis, and provide more opportunities to answer other research questions.

Regarding the statistical analysis we used in question 3, if we were to do this test again, it would make more sense to find a difference in mean satisfaction rating for the different groups, rather than a difference in proportions. This is because dividing "Satisfied" and "Unsatisfied" into binary groups based off being above or below a 50 satisfaction rating is more subjective, and this threshold can be changed. It would be more effective to find the mean ratings from the raw data as this arbitrary threshold might not be representative of those who were actually satisfied or unsatisfied. In addition, by creating these binary satisfied/unsatisfied 'bins', there were very few respondents that were unsatisfied, meaning the actual number of possible differences in proportion in the null distribution was very small, making it likely we would fail to reject. In the end, the difference in proportions is still functional and provides an interesting result, but in the future the satisfaction threshold could be played with or a difference in means could be used to see if there are different findings. During sampling, we could have also asked a question in our survey directly whether somebody was satisfied or unsatisfied with their housing (binary option: yes or no). This could have removed some of the subjective analysis of determining satisfaction based on the number scale.

One of the limitations with our project is that we decided to lump together everyone who wanted a room bigger than a double into the suite category. If we were continue this project,

we might expand our question and explore the proportion of people who wanted each kind of suite (triple, quad, five-person, six-person). In general students are relatively happy with their sophomore housing and the housing selection process in general. Middlebury College changed the selection process this year. Last year, there were 2 different lottery draws, one for people who wanted a suite and one for those who wanted doubles or singles. If you didn't get a suite in the suite draw, then you were automatically entered in the double/single draw. We are curious if satisfaction with the system this year has increased since last year. This would show us if the college made the good decision changing the system. If we had access to satisfaction data from last year, we could perform a hypothesis test on the difference in means to see if satisfaction has increased this year.