Exploring the Influence of Geographic Origin for Freshman Battell Residents

Julia Randolph and Tessa Herbst

2024-12-15

```
## Rows: 30 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): Timestamp, Hometown, Mountain, Know_creemee, New_england
## dbl (7): VT_visits, Ski_age, Ski_level, Like_creemee, Avg_snow, Avg_temp, Di...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Introduction

For this project, we researched four questions regarding the influences of geographic origin, with freshman students who live in Battell being our target population for every question.

Question 1: Is the average comfort level of New Englanders with skiing greater than the average comfort level with skiing of non-New Englanders?

Question 2: Is the proportion of New Englanders who had heard of a Creemee before they arrived at Middlebury greater than the proportion of non-New Englanders who had heard of a Creemee?

Question 3: What percentage of Middlebury students who live in Battell had been to Vermont before coming to/visiting Middlebury for admissions reasons?

Question 4: Does being within a 3.5-hour drive from a ski mountain affect the age participants first skied?

We both are from different parts of the country, and we see that we have different relationships to Vermont. We are curious to explore this on a larger level with the freshmen in Battell.

Data Collection

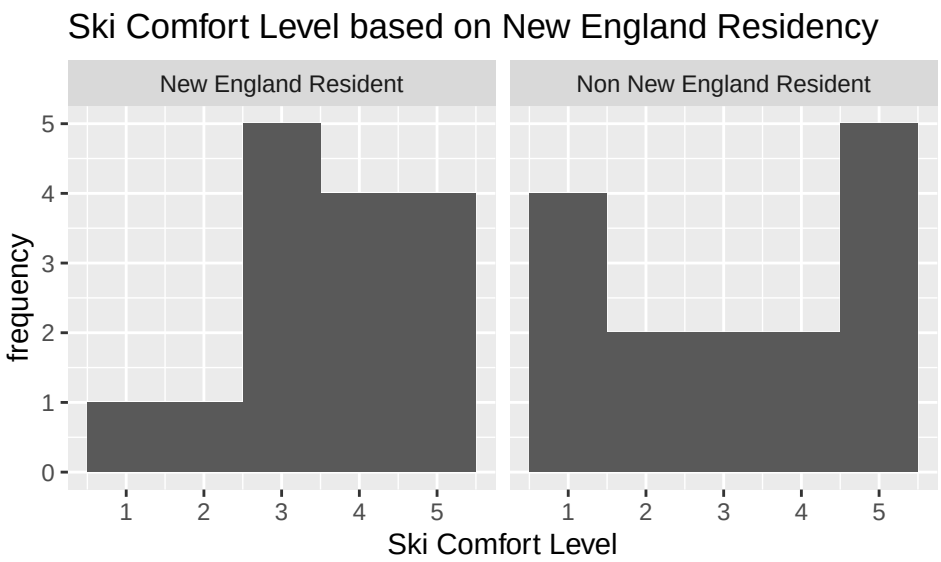
Using R, we randomly sampled 35 dorm rooms in Battell. We then distributed surveys to the selected rooms in the form of a QR code. Although we tried to conduct a random sample, there is nonresponse bias in the sample, as participants had the choice of whether to participate or not; many chose not to participate.

Variables:

- Hometown: where the participant lives
- VT_visits: number of times non-residents had been to Vermont before coming to Middlebury
- Ski_age: age the participant first skied
- Ski_level: comfort level with skiing (1-5)

- Mountain: does the participant live within a 3.5-hour drive of a ski mountain (yes/no)
- Know_creemee: had the participant heard of a creemee before Middlebury (yes/no)
- Like_creemee: how much the participant likes creemees (1-5)
- Avg_snow: average annual snowfall in the participant's hometown
- Avg_temp: average winter temperature in the participant's hometown
- Dist_midd: distance from the participant's hometown to Middlebury College
- New_england: does the participant live in New England (yes/no)

Methods: Research Question 1:



The above histogram compares the ski comfort level for New England and non New England residents, with the lowest comfort level being 1 and the highest being 5. The histogram does not display an obvious difference in the mean ski comfort level of New England residents compared to non-New England residents. Based on the original data, the observed difference in means (mean for New England Residents - mean for non-New England Residents) is 0.4666667. In order to find the true difference in means for the target population of freshmen in Battell, we performed a hypothesis test.

As stated above, for this question we conducted a hypothesis test with difference of means to see if the average skiing comfort level of New England residents is greater than the comfort level of Non-New England residents, with the target population being freshmen who live in Battell. We can use CLT because the conditions are met. In other words, as shown in the histogram, the distribution of ski comfort levels is approximately normal as there are no clear outliers, and there is also independence as a random sample was conducted.

We used the New_England variable, our explanatory variable, which listed if the participant was a New England resident or not. This was a mutated column we made depending on their hometown. We then made new data frames, one of which only included New England Residents and the other included only non-New England Residents. We then pulled the Ski_level variable as the response variable for this research question.

H_0 : The mean skiing comfort levels are the same between New England Residents and Non New England Residents.

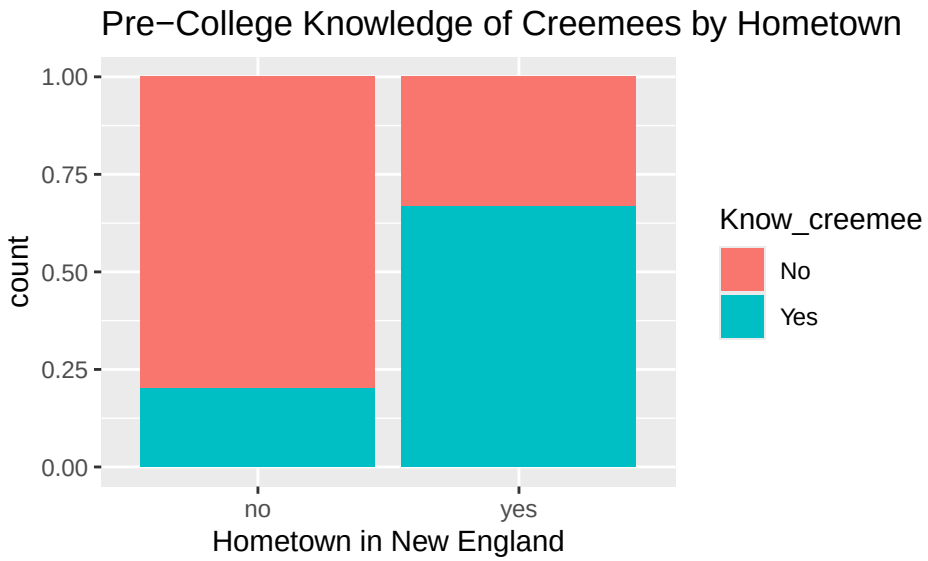
H_A : The mean skiing comfort levels are not the same between New England Residents and non New England Residents.

The significance level for this question is set at 0.1.

Results: Research Question 1:

For our hypothesis test, our p-value is 0.1973957, which is larger than our significance level, so we fail to reject the null hypothesis. Therefore, the data does not provide convincing evidence that the mean ski comfort level of New England Residents is greater than that of non-New England Residents.

Methods: Research Question 2:



The observed difference in proportion is 0.4666667; as seen in the above bar chart, the data displays a rather significant difference, with a much greater proportion of New Englanders having heard of creemees compared to non-New Englanders. Because of this, it would make sense to test the hypothesis that there is truly a difference between the proportion of New Englanders who'd heard of a creemee before Middlebury and the proportion of non-New Englanders who'd heard of a creemee before Middlebury. Coincidentally, 15 of our observations live in New England and 15 do not, so the sample size of each group is 15.

The variables we used to answer this research question were Know_creemee and New_england; New_england was our explanatory variable, and Know_creemee was the response variable that we were testing proportions for. Our population of interest was all freshmen students who live in Battell.

For this question, we used simulated randomization for two proportions to test the hypothesis that a higher proportion of New Englanders (P1) had previously heard of a creemee than the proportion of non-New Englanders (P2) that had previously heard of a creemee. We used simulation rather than CLT because the the success-failure condition for CLT was not met.

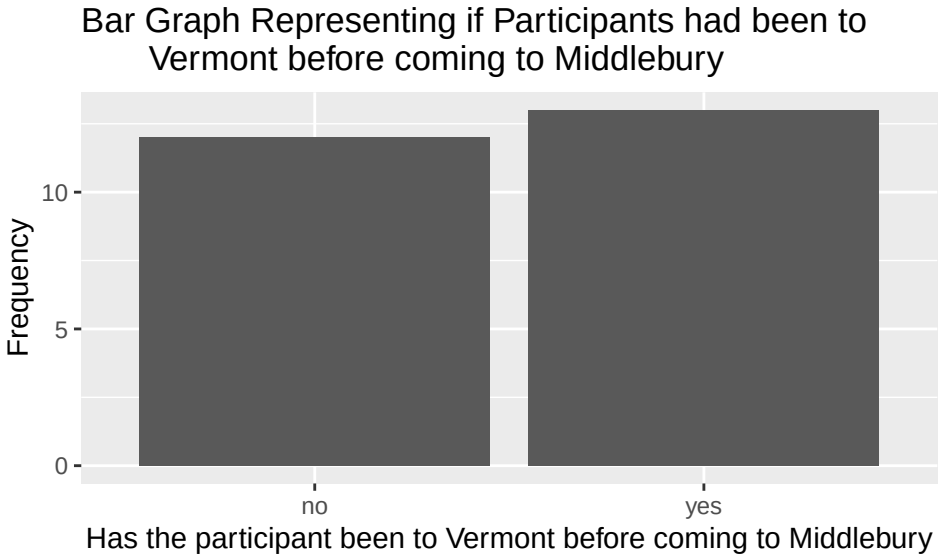
H_0: P1 - P2 = 0 H_A: P1 - P2 > 0

The significance level for our analysis is 0.05.

Results: Research Question 2:

The p-value of 0.01 is less than the significance level, so we reject H_0. The data provide convincing evidence that the proportion of New Englanders who had heard of a creemee before Middlebury is greater than the proportion of non-New Englanders who had heard of a creemee.

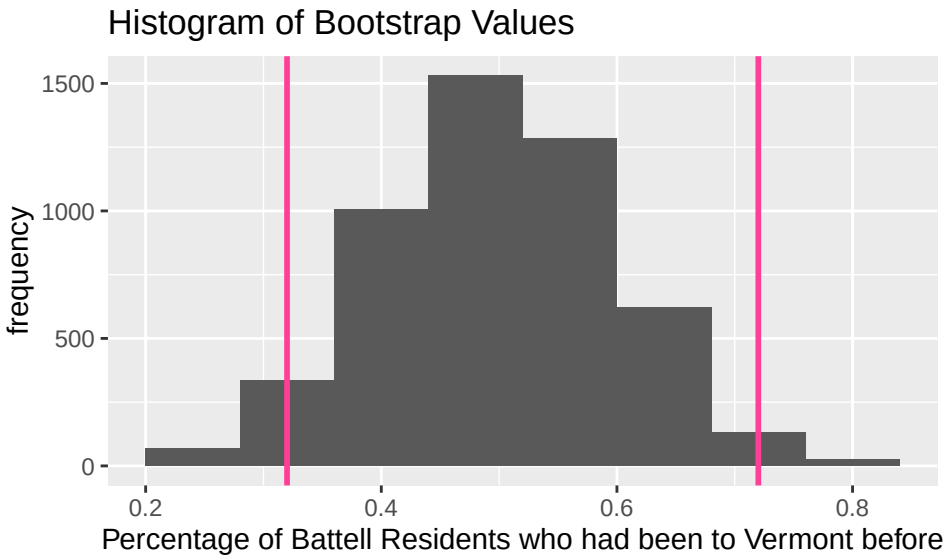
Methods: Research Question 3:



First, we made a bar graph representing if the participants in our sample had been to Vermont before coming to Middlebury (excluding for admissions purposes) and we noticed that there seemed to be a few more observed individuals who had been to Vermont before as opposed to those who had not, sparking our interest in finding the true percentage of students who had been to Vermont before coming to Middlebury

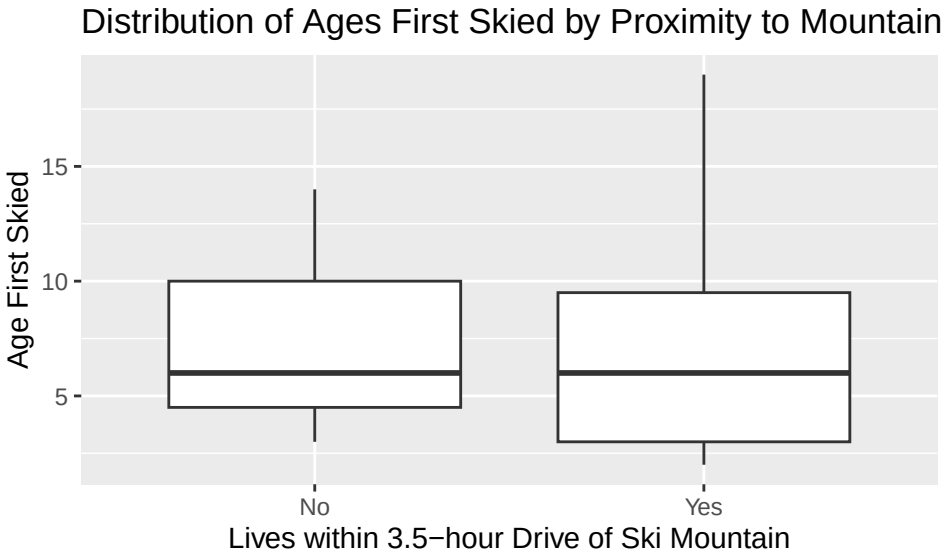
We then used a simulation scheme to find a bootstrap value for this question. We created a new data table in which the mutated VT_visits data column says "yes" if the value was greater than zero (the participant had been to Vermont) and "no" if the value was zero (the participant had never been to Vermont). As we were only looking into one variable, VT_visits was the only variable used for this question. We also filtered out all the NA results which represented participants who were full time Vermont residents, which meant that our data frame had 25 observations instead of 30.

Results: Research Question 3:



We are 95% confident that the true proportion of Middlebury students who live in Battell that had been to Vermont before coming to Middlebury (excluding for admissions purposes) is between 0.32 and 0.72. The Pink Line Represents the lower and upper bounds respectively. 5000 iterations of the Bootstrap were done.

Methods: Research Question 4:



The above boxplot suggests that there is not a significant difference in the age that participants first skied between those that live within a 3.5-hour drive of a ski mountain and those that do not. In fact, the two groups have the same median of age first skied, which is 6. However, we were interested in confirming if this is the case, so we still tested for a relationship, using simulation-based testing to create a sampling distribution.

As we are looking at the relationship between two variables, the statistical method we used was single linear regression, with one variable, Mountain, being categorical. We then conducted simulation-based hypothesis

testing to see if living within 3.5 hours of a ski mountain affects the age that students in Battell first skied.

For this research question, we used the variables Ski_age and Mountain. Our population of interest remains all freshmen students who live in Battell, but we filtered the data set to remove observations where Ski_age = NA, which resulted in a sample size of 26.

We intend to fit a linear regression model that takes the form of the following: Ski_age = $\beta_0 + \beta_1$ * Mountain

As there are only two levels of the categorical variables, the linearity condition is satisfied. On the other hand, although unlikely, it is possible that independence is violated due to different observations being related to each other, and knowing one could have an impact on the other. For example, if two participants have been friends with each other before Middlebury, their responses may be similar due to that relationship rather than due to living near a ski mountain; if one observation lives within a 3.5-hour drive of a ski mountain and has been close friends with another participant who lives farther away from a mountain, that friend may be more likely to have gone skiing at a younger age than they would have because they had the opportunity to ski with their friend. However, because the data do not appear to be dependent on each other, and they were collected in a random sample, this is highly unlikely and we are assuming independence.

H_0: $\beta_1 = 0$ H_A: $\beta_1 \neq 0$

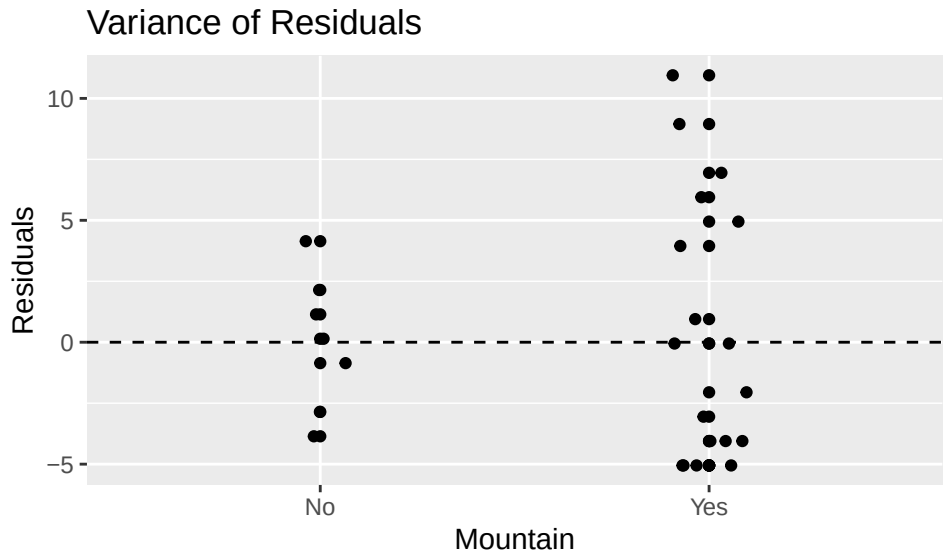
Let β_1 = the slope of the true linear regression model; the true amount of years the age students first skied changes by if they live within 3.5 hours of a ski mountain.

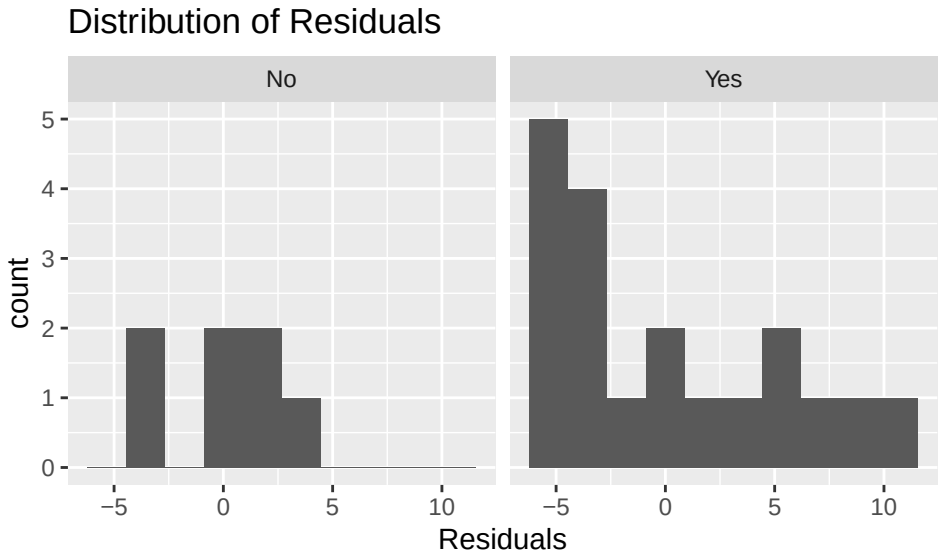
The significance level for our analysis is 0.05.

Results: Research Question 4:

The fitted model, as the result of linear regression, for our data is as follows:

Ski_age = 7.4285714 + 0.0451128 * Mountain





The condition of normal residuals does not seem to be completely met; because of this, the linear regression model likely is not a representative fit. As this is just a preliminary model, for now we are assuming that the normal residuals condition is met.

Additionally, the condition of equal variance is not met either. Based on the scatterplot of residuals, there does not seem to be equal spread between observations that live within 3.5 hours of a ski mountain and those that do not. The residuals of observations who live near a mountain are significantly more spread out than those that do not, ranging from -5 to values greater than 10. In order for all the observations to be visible, including repeated Ski_age data, we wrangled the plot to show the repeated points side-by-side. Again, this occurs because the model is just preliminary; as our findings show no significant relationship, it makes sense that the conditions for fitting a linear model may not be met.

The p-value is 0.958, which is significantly greater than the significance level, so we fail to reject H_0. The data do not provide sufficient evidence that living within a 3.5 hour drive of a ski mountain affects the age that freshmen in Battell first skied. "'

Discussion

In conclusion, our most interesting finding was that a Battell resident's hometown seems to affect their knowledge of a creemee existing. With a p-value of 0.01, we rejected the null hypothesis, concluding that the data provides convincing evidence that the proportion of New Englanders who had heard of a creemee before coming to Middlebury is greater than the proportion of non-new Englanders who had heard of a creemee before Middlebury. We did not see any connection between New England and Non-New England Residency affecting one's ski confidence or the distance to a mountain affecting the age first skied; both of these hypothesis tests resulted in us failing to reject the null given high p-values. For our first research question, we concluded, with 95% confidence, that the true proportion of Middlebury students who live in Battell that had been to Vermont before coming to Middlebury (excluding for admissions purposes) is between 0.32 and 0.72. This is a rather wide interval, so our results were not very conclusive about the true percentage; however, we can interpret this confidence interval to deduce that around half of the freshman population in Battell, who do not live in Vermont, had traveled to Vermont before coming to Middlebury.

Although we did our best to create a random, representative sample of freshmen who live in Battell, it is likely that we did not fully achieve that, which could have introduced Type 1 or Type 2 errors. Since participants could choose whether or not to take the survey, there is nonresponse bias in the sample. Additionally, it is possible that some observations are not independent of each other; for example, if two participants are friends, they could have responses that are dependent on each other; if one friend in New England had heard

of a creemee, they couuld have told a friend who lives outside of New England about it. We do feel that this is incredibly unlikely, however. Moreover, we can not be completely confident about the data collected because many of the survey questions are subjective or dependent on memory which would be difficult to confirm.

In order to have a more conclusive analysis of the data, we could have made better use of exploratory data analysis to refine our research questions and statistical methods; many of our findings lacked interesting or new information, and this was made clear after simple visualizations of the data. The statistical analysis we then conducted on these questions seemed rather redundant as it was obvious from the first step that there would be no findings of interest. For example, in Research Question 4, it is very clear from the initial boxplot that there is no linear relationship between proximity to a ski mountain and the age students first skied; because we looked past this, we fit a linear regression model to conduct a hypothesis test that resulted in a model that violated both the normality and equal variance conditions of an appropriate fit, and did not result in much significant information. In this case, it may have been better to conduct a test of the difference in means, rather than attempting to force a linear model onto the data.

If we were to continue work on the project, we would look into additional research questions that seem relevant after exploring the data. After a quick glance at the data, it seems as though there may be other relationships of interest, such as whether living in New England affects the number of visits a student in Battell has taken to Vermont, or how average annual snowfall is related to the participant's comfort level with skiing. Additionally, if we were to start over with the project, we would work to ensure that nonresponse bias was less prevalent; one way we could have done this is by speaking to the students in each room face to face to collect the data, which would increase the incentive to participate and hopefully decrease the amount of nonresponse bias. Additionally, we would increase our sample size, potentially to include all freshman dorms or more students living in Battell. By increasing the size of our study, we could also potentially conduct a hypothesis test on just New England Residents or only non-New England Residents, or expand our population of interest to be all freshman students at Middlebury.