Title: An Analysis of Survey Data on Students from the University Park Campus at Penn State

Authors: Jonathan Fivelsdal, James Fulton, and Will Goldenberg

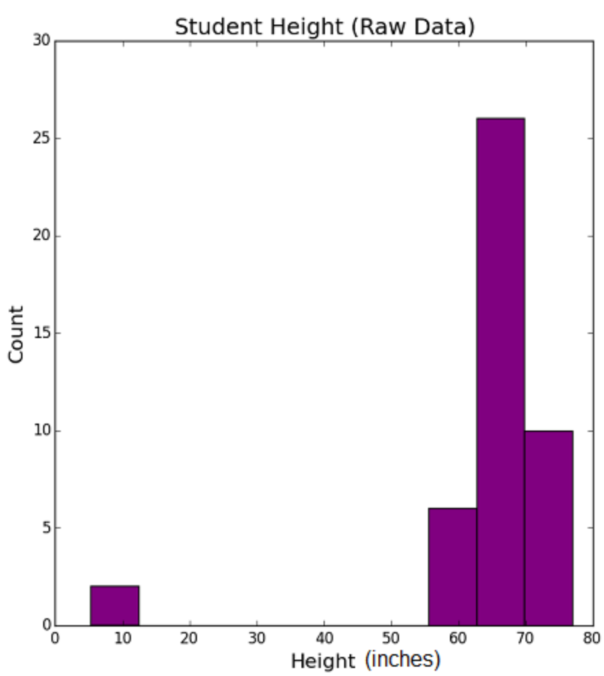
Course: Applied Statistics – STAT 500

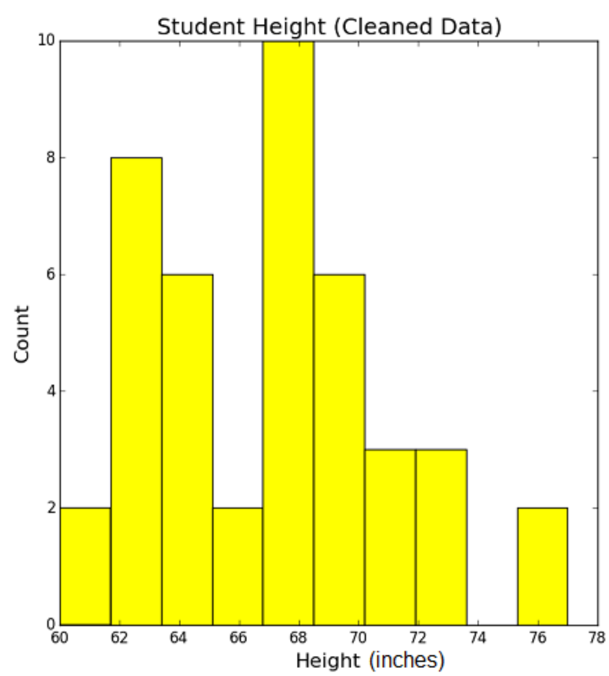
Professor: Dr. Basak

Title: An Analysis of Survey Data on Students from the University Park Campus at Penn State

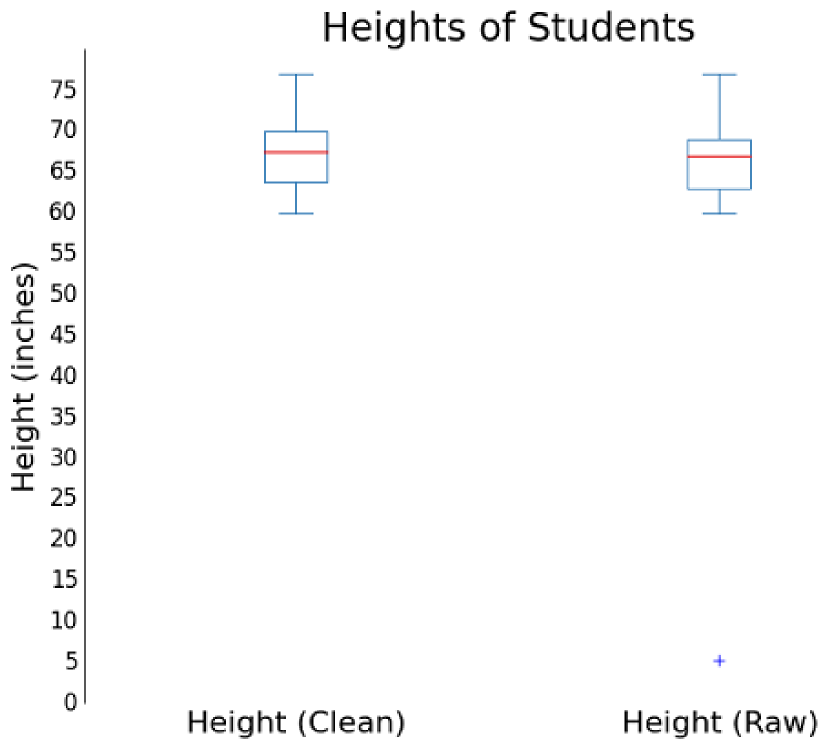
Introduction: For years college and university related data has been used for various purposes and probably most notably for the purpose of ranking colleges and for colleges to present statistics on their incoming freshman class on their website. Typically, the student data that is presented in college ranking publications or on college websites is data on the test scores and demographic data. What is interesting about the data that we consider in this report is that some of the data we consider is not typically presented in college ranking publications or on the website of a college. For example, college ranking publications or college websites typically do not provide information on how much television a student watches or the body measurements of the students that attend a college. Although the data we consider in this report does not consider all of the students that attend a particular college or even an entire class year of students, statistically significant relationships have been obtained by examining the properties in the class data that is considered in this report. We examine survey data collected on 44 Pennsylvania State students from a class offered at the University Park campus. Questions were asked about the students gender, race, the college in which their field of study is in, their natural hair color, if they have dyed their hair or not, the number of hours they spend studying, the number of hours they spend watching television, the GPA they have in college, what the student would consider to be an ideal weight and ideal height, and the students were also asked about various body measurements.

Exploratory Data Analysis: Before trying to find insights within the data, we performed some exploratory data analysis in order to identify and correct errors within the data. One of the variables in which errors were present is the student height variable. The survey that was given to each student asked for their height in inches. After creating a histogram of the raw data, we observed that there was data that suggested there were students that had height measurements below 24 inches (2 feet).



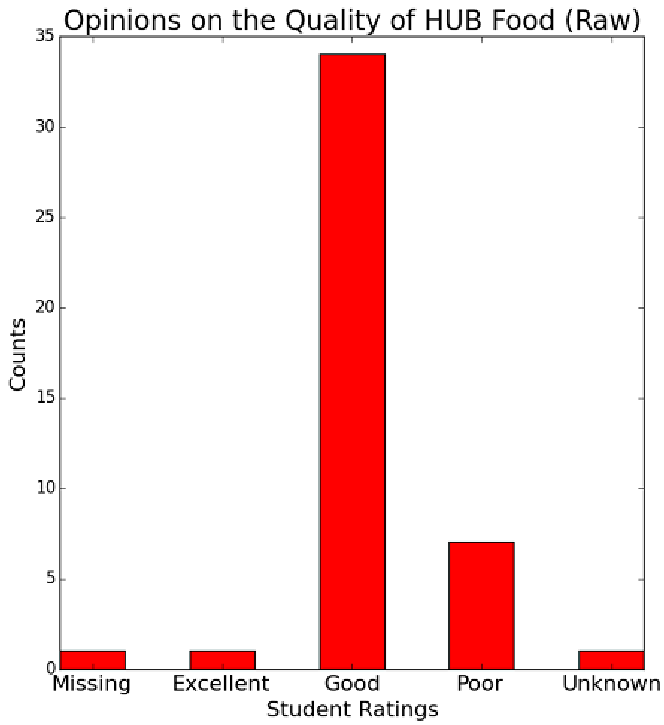


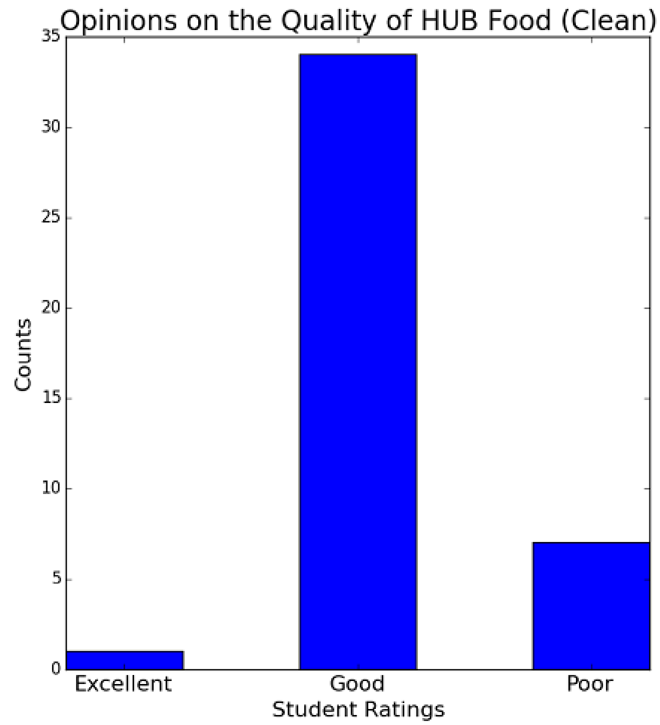
We also created a box plot to examine the student height data.



The boxplot shows that outliers (represented by the cross mark) at approximately a height of 5 inches. We believe that the students that recorded their height to be about 5 inches was confused and thought that the survey was asking for their height in feet instead of inches. We removed the two height measurements of 5.3 inches from the data. After cleaning the height variable (which is the data in the yellow histogram and the left side of the boxplot), the heights of students ranged from about 60 inches (5 feet) to 77 inches (6.42 feet).

Another variable we examined for errors is the variable HUB food. This variable asked students about their opinion of the food served at the campus cafeteria (the HUB). In order to identify problems with this variable, we created a bar graph that showed the number of students that gave a rating for each valid category. The three choices of excellent, good and poor were presented on the survey as valid ratings for their opinion of HUB food. Out of the 44 students surveyed, there were 42 responses that were one of the 3 choices of excellent, good, and poor. There was one missing response for the question on the quality of food served at the HUB and there was one invalid response of “Unknown” for the rating of the quality of food served at the HUB. We provide a figure that shows the raw data with the error and missing value along with the cleaned version of the data on opinions of HUB food without the missing rating and the invalid “Unknown” value.





After removing the missing data and erroneous value, we ended up with valid responses from 42 students which are depicted by the blue bar graph on the right. In the clean form of the data there is one student that rated the food as excellent, there are 34 students that rated the food to be of good quality, and there are 7 students that rated the food to be of poor quality.

Analysis and Results: After performing some exploratory data analysis and cleaning up data, we performed some t-tests to identify relationships within the data set. Later we use the Chi-square test to examine the relationships between a pair of categorical variables.

T-Test Hypotheses for Comparing Mean Height for Caucasian females and Asian females

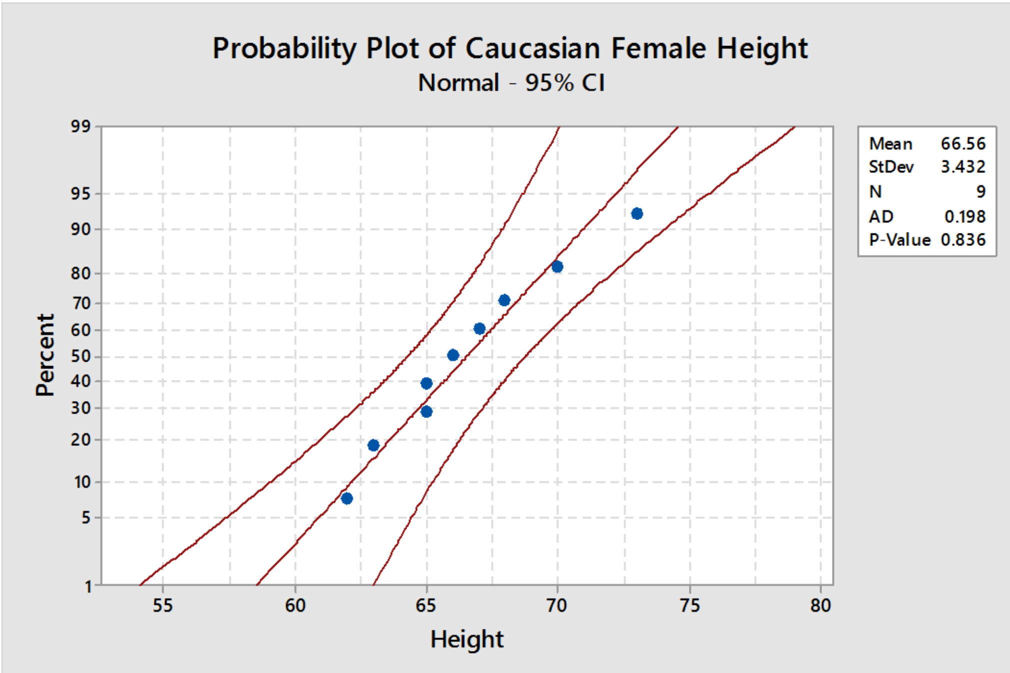
Our initial t-test involves the gender and height data. The t-test is used to determine if the height for Caucasian female students is greater than the mean height for Asian female students. Let μ_{CFH} represent the mean height for Caucasian female students and let μ_{AFH} represent the mean height for Asian female students. The hypotheses for the first t-test are as follows:

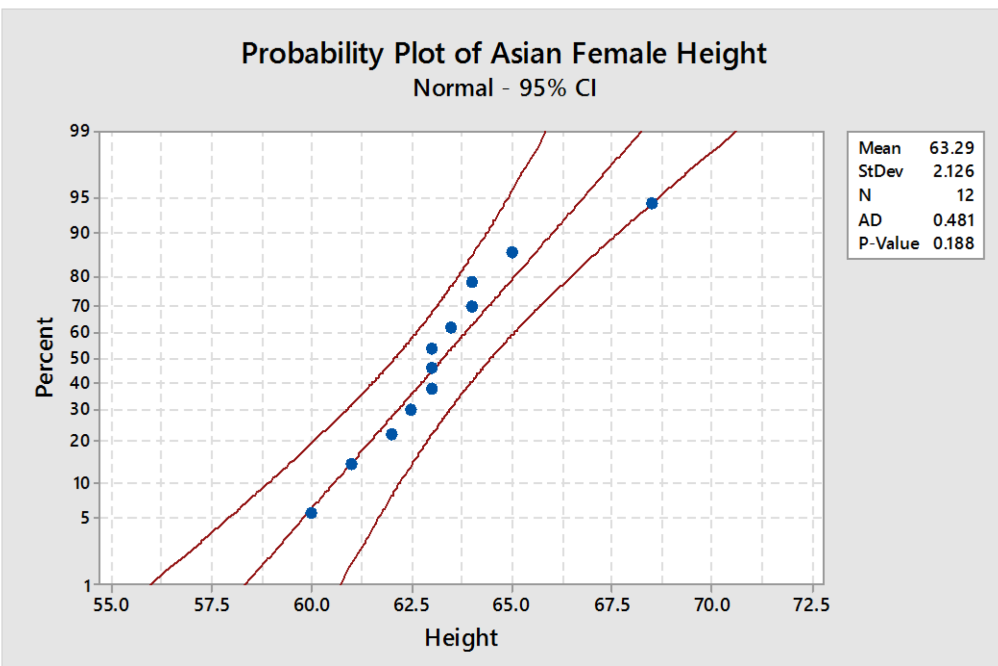
H_0: $\mu_{CFH} = \mu_{AFH}$

H_a: $\mu_{CFH} > \mu_{AFH}$

Since the sample size of height measurements for Asian female students is less than 30 and the sample size of height measurements for Caucasian female students is less than 30, we proceeded to check if the normality condition held in so that we could perform the t-test.

We have provided a probability plot of the height data for Caucasian female students and the height data for Asian female students.





The data points in the plot for the heights of Caucasian female students lie well within the confidence bands of the probability plot and thus we believe that the data in that sample is approximately normal. Similarly, the data points in the plot for the heights of Asian female students lie well within the confidence bands of the probability plot and thus we believe that that sample of data is approximately normally distributed. Now that we have checked the normality of the two samples we proceed to test the relationship between the two samples of height data.

T-Test Results Comparing Mean Height for Caucasian females and Asian females

There are 12 Asian females in the sample and there are 9 Caucasian females in the sample. After performing Levene's test we obtain a test statistic of 2.05 and a p-value of 0.168. From the results of Levene's test it seems plausible to use the pooled variance test. We perform an F-test at the 95% confidence level and get a test statistic of 0.3838, numerator degrees of freedom of 11, and denominator degrees of freedom of 8 and a p-value of 0.1434. Since the p-value is greater than 0.05, we fail to reject the null hypothesis and thus at the 95% confidence level there is evidence to suggest that the variance of the heights of Caucasian female students is equal to the variance of the heights of Asian female students. Given the results of Levene's test and F-test, we will perform a 2-sample t-test using a pooled variance.

After performing the t-test we obtain a t-statistic of 2.689, a p-value of 0.007, and we have degrees of freedom of 19. Since the p-value is less than 0.01, we reject the null hypothesis at the 99% confidence level. There is enough evidence at the 99% confidence level to suggest that the mean height of Caucasian females is greater than the mean height of Asian females.

Originally there was height survey data for 13 Asian females. However, one of the recorded values was 5.3 for height. The survey asked for each student to enter in their height in inches and a height of 5.3 inches is not realistic for a human and is thus considered an error. The student that answered with this value of 5.3 was not included in the t-test performed above.

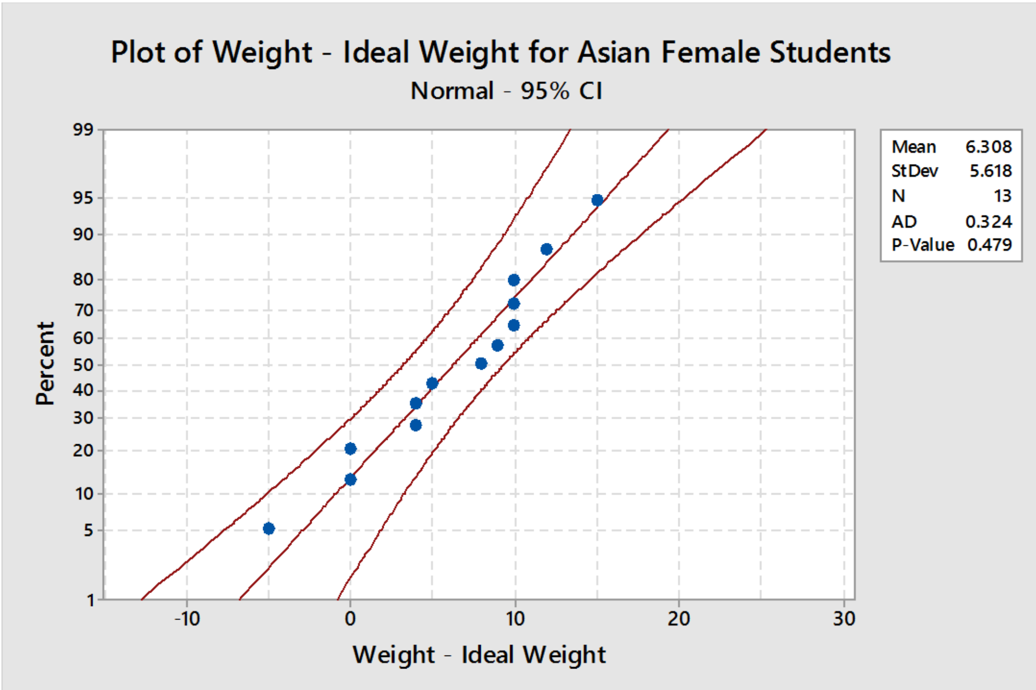
T-Test Hypotheses Comparing Mean Weight and Mean Ideal Weight for Asian Female Students

Next we conducted a paired t-test at the 98.5% confidence level (α=0.015) to test if the mean actual weight of Asian female students was different from the mean ideal weight of Asian female students. With the mean actual weight of Asian female students being represented by μ_{AFW} and the mean ideal weight of Asian female students being represented by μ_{AFIW}, we then have the following two hypotheses for our paired test:

H_0: $\mu_{AFW} = \mu_{AFIW}$

H_a: $\mu_{AFW} \neq \mu_{AFIW}$

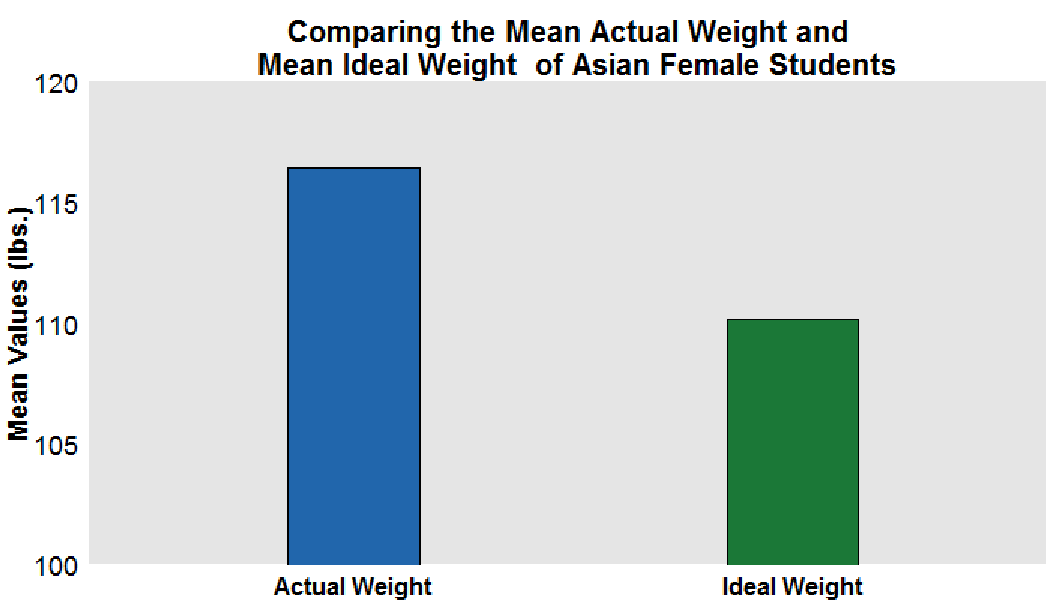
Due to the small sample size, conditions of normality were checked before performing the t-test. Normality conditions were checked by creating a probability plot of the difference between the actual weight for Asian female students and the ideal weight of Asian female students.



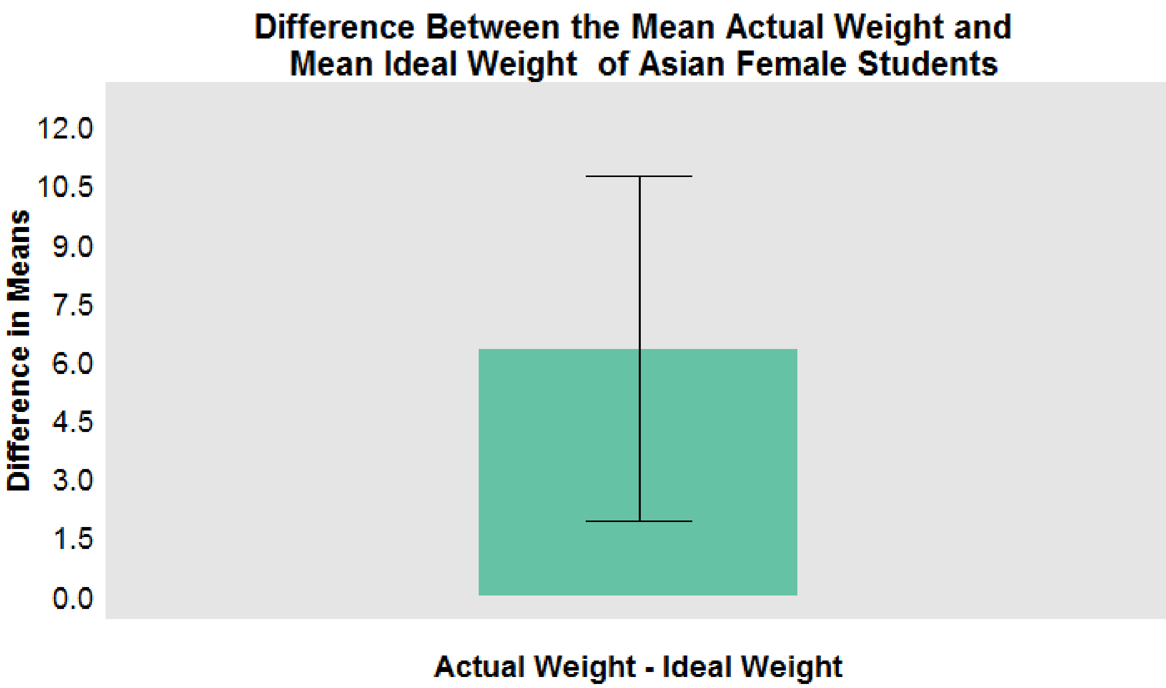
The data points in the probability plot for the difference between weight and ideal weight for Asian female students lie well within the confidence bands of the probability plot and thus we believe that the data in this sample is approximately normal.

<u>T-Test Results Comparing Mean Weight and Mean Ideal Weight for Asian Female Students</u>

There are 13 Asian female students in both the sample for actual weight values and ideal weight values. We provide a plot of the sample means for both the ideal weight of Asian female students and the actual weight of Asian female students.



The sample mean actual weight for Asian female students is 116.46 pounds and the sample mean ideal weight is 110.15 pounds. We used the weight data to perform a two-tailed, paired t-test to determine whether or not the population mean for the ideal weight of Asian female students is different from the population mean for the actual weight of Asian female students. The difference between the mean actual weight for Asian female students and the mean ideal weight for Asian female students is 6.3 pounds. Below is a graph showing the difference between mean actual weight for Asian female students in the sample and the mean ideal weight for Asian female students in the sample along with a corresponding confidence interval.



The 98.5% confidence interval we obtained estimates the true difference in the actual weight of Asian female students and the ideal weight of Asian female students to be between 1.89 pounds and 10.73 pounds (the error bar depicts the confidence interval in the graph above). Other values we obtained from performing the t-test are a t-statistic of 4.048, a p-value of 0.0016, and 12 degrees of freedom. At a significance level of 0.015 (α=0.015), we reject the null hypothesis since our p-value is below the pre-set significance level. We similarly reject the null hypothesis based on the confidence interval since the confidence interval does not include the value of 0 which would indicate that there was no difference between the mean of the actual weight for Asian female students and the mean ideal weight for Asian female students. Therefore we have evidence at the 98.5% confidence level to suggest that the mean weight for Asian female students is different from the mean ideal weight for Asian female students.

Power Analysis: Detecting a One Pound Difference Between Ideal Weight and Actual Weight

We now will perform power calculations in order to determine the necessary sample size for a hypothesis test that will test the research hypothesis that graduate students have a lower mean ideal weight than their mean actual weight and this test will be able to detect a difference of one pound at the 80% power level. The actual weight and ideal weight of the graduate students are paired data. We will use a paired t-test. We provide a table of basic statistics for both the actual weight of graduate students and the ideal weight of graduate students.

	N	Mean	StDev	SE Mean
Weight	43	144.28	29.05	4.43
Ideal-weight	43	136.35	27.98	4.27
Difference	43	7.93	12.82	1.96

The standard deviation of the difference of the groups is 12.82. Let μ_{GW} represent the mean weight of all graduate students and let μ_{GIW} represent the mean ideal weight of all graduate students. We would like to perform power calculations for the following one-sided paired t-test:

H_0: $\mu_{GIW} = \mu_{GW}$

H_a: $\mu_{GIW} < \mu_{GW}$

We perform a sample size calculation in order to find out how many students we would need to sample to obtain 80% power in order to achieve a minimum detectable difference of 1 lbs. at the 0.05 alpha level.

This output from Minitab displaying the results of the sample size calculations:

```
Calculating power for mean paired difference = difference
α = 0.05  Assumed standard deviation of paired differences = 12.82


            Sample  Target
Difference    Size   Power  Actual Power
        -1    1018     0.8      0.800181
```

The power calculations for the left-tailed, two-sample test indicate that we need 1018 pairs of graduate students in order to detect a 1 pound difference between the mean actual weight and mean ideal weight among graduate students by performing a t-test with 80% power.

Examining the Relationship Between the Gender of a Student and Hair Dyeing

The two qualitative variables we will create hypotheses about are the race variable and the ‘hair dyed’ variable. Next we provide a table of observed values and expected values for the number of people that chose whether or not to dye their hair by gender. We have provided the expected values in parentheses.

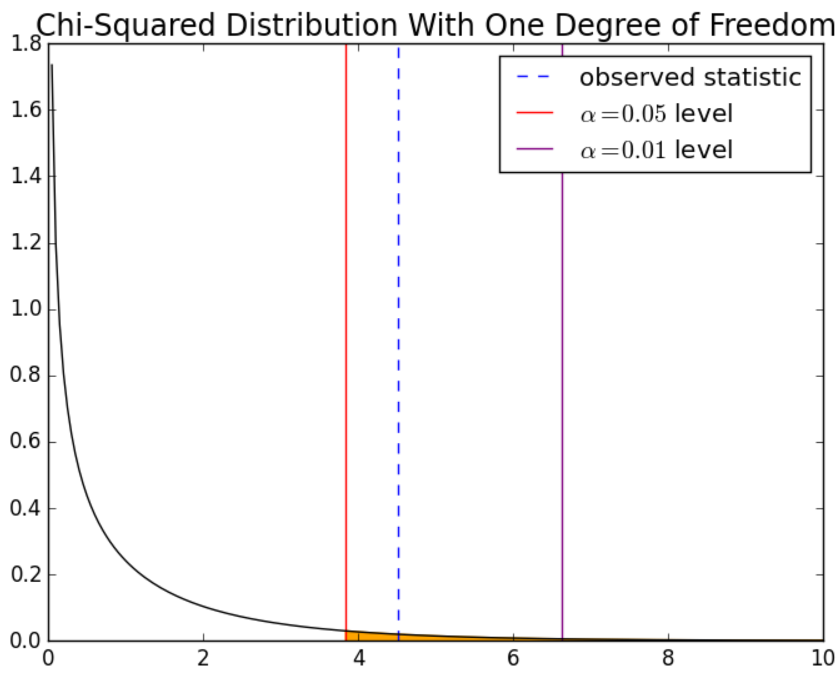
Gender	No (Natural)	Yes (Dye Hair)	Row Totals
Female	13 (16.28)	12 (8.72)	25
Male	15 (11.72)	3 (6.28)	18
Column Totals	28	15	43

From the table above we see that less than 20% of the expect counts are less than five and so we can use the chi-square test. Since we have a 2 x 2 table we obtain one degree of freedom. We will perform the hypothesis test using the following null hypothesis and alternative hypothesis:

H_0: The decision for a student to dye their hair is independent of their gender.

H_a: The decision for a student to dye their hair depends on their gender.

After performing a chi-square test, we get a chi-square statistic of 4.52. Below we have a figure which contains the chi-squared distribution with one degree of freedom, the critical chi-square statistic at the 0.05 significance level, the critical chi-square statistic at the 0.01 significance level, and the chi-square test statistic that we observed from performing the chi-square test.



The red line corresponds to the critical chi-square statistic at the 0.05 significance level which is approximately a value of 3.84. The dotted line represents the test statistic that we obtained from the chi-square test which is 4.52. The orange color under the chi-square curve represents the rejection region when the chi-square test is conducted at the 0.05 significance level. Since we tested our research hypothesis at the 0.05 significance level and our observed test statistic is greater than the critical value at the 0.05 significance level, we reject the null hypothesis. Notice that if we conducted the chi-square test at the 0.01 significance level, we would fail to reject the null hypothesis (we would need a chi-square statistic greater than 6.63 in order to reject the null hypothesis at the 0.01 significance level).

From the chi-square test we obtain a p-value of 0.033. Similar to the rejection region approach we used, we reject the null hypothesis at the 95% confidence level since $0.033 < 0.05$. Therefore at the 95% confidence level, there is evidence to suggest that the student's choice as to whether or not to dye their hair depends on gender.

Summary and Future Work: Select variables measured on 44 students at a class at the University Park campus at Penn State were examined. The first variable we examined was student height. We created a box plot and a histogram in order to observe the spread of the height data and if there were any errors in the height data. After plotting the height data of the students we saw that there were height measurements of less than one foot that looked suspicious. There were in fact two student heights that were recorded as 5.3 inches. These points were removed and we plotted the data with the other 42 student height measurements. After plotting the clean student data, the student heights ranged from 5 feet to about 6 foot 2 inches. The other variable we examined was the HUB food variable which measured student opinion of the food at the cafeteria at Penn State. Again we plotted the raw data and clean data for the HUB food variable but this time we used a bar graph in order to look for errors. The raw data included one value of "Unknown" which was an error. In the raw data, there was also one missing value. After removing the erroneous value of "Unknown", we plotted the clean data without the error and missing value which showed that there were 34 students that rated the food goo, 7 students rated the food poor, and 1 student rated the food excellent.

After doing some exploratory data analysis, we conducted a few hypothesis tests. The first hypothesis test we performed compared the heights of Caucasian female students with the heights of Asian female students. The research hypothesis states that the mean height of female Caucasian students is greater than the mean height of Asian students. From the independent t-test we performed, we obtained a p-value of 0.07 and thus found that there is evidence to suggest that the mean height of female Caucasian students is greater than the mean height of Asian students. The second hypothesis test we performed tested the research hypothesis that the mean weight of Asian female students is different from the mean ideal weight of Asian female students. From the test we obtained a p-value of 0.0016 and thus we rejected the null hypothesis at the 98.5% confidence level. Therefore there was evidence at the 0.015 significance level that the mean weight of Asian female students is not equal to the mean ideal weight of Asian female students. After comparing the mean weight and mean ideal weight of Asian female students, we performed a power calculation to determine how many graduate students we would need in order to test if the mean ideal weight of graduate students is less than the mean weight of graduate students such that we could detect a minimum difference in weight of one pound and conduct the test at 80% power. The sample size calculations we made were for a left-tailed paired test and the result was that both the mean weight and mean ideal weight samples would need to have a sample size of 1,018. The last hypothesis test we conducted tested the research hypothesis that the decision for a student to dye their hair depends on their gender. In order to test this research hypothesis we used a chi-square test at the 0.05 significance level. From the chi-square test, we obtained a p-value of 0.033 and so there is evidence at the 0.05 significance level that suggests that the decision for a student to dye their hair depends on their gender.

Although we found some interesting relationships within the student survey data, our analysis is not exhaustive. Variables that we did not consider in this report were what sport a student plays, hair color, number of pairs of jeans, GPA, hours of studying per week, hours of television watched per week, and different body measurements (measurements of the left wrist, right wrist, and the head circumference of a single student measured by two different students).Additional work could be carried out in the future could include investigating whether or not the opinion of the food served at the HUB depends on what sport a student plays, test if the mean time spent watching television per week is greater than the mean time spent studying, and test if there is a relationship between gender and the number of pairs of jeans owned by a particular student.