Deliverable\_One

February 22nd, 2021

Nathaniel Jones

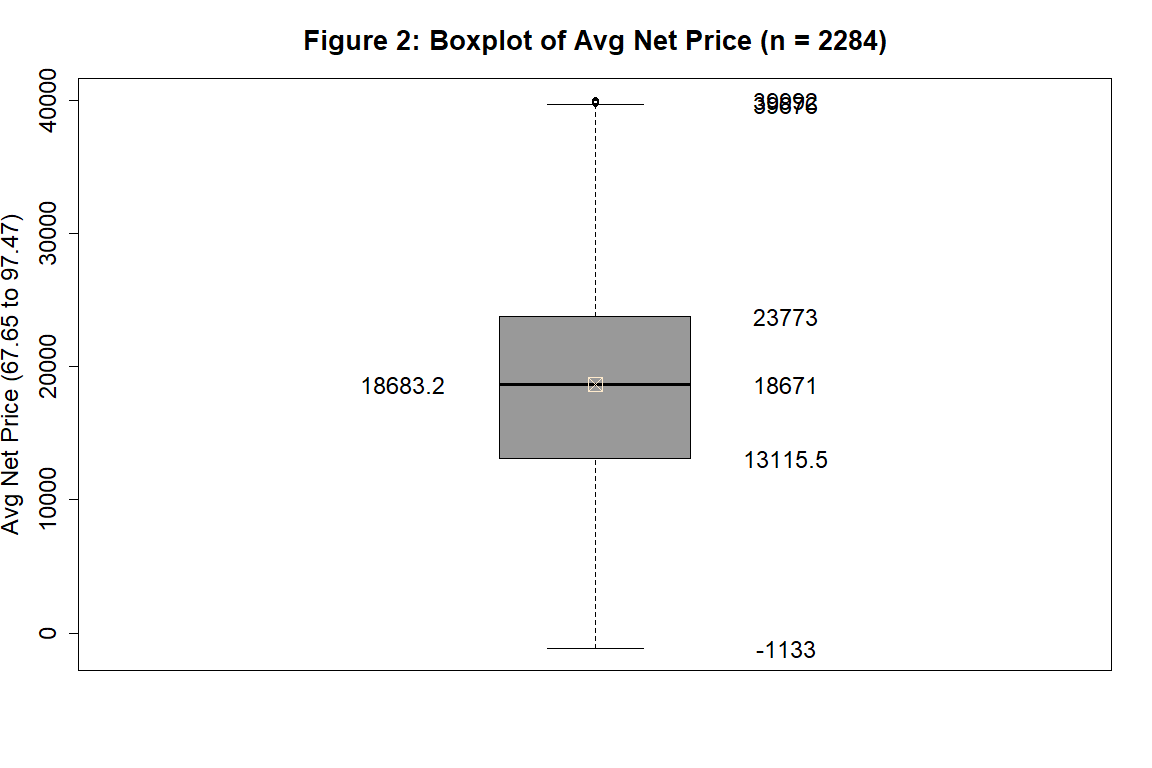
working %>% select(NPT4\_ALL) %>% summary()

## NPT4\_ALL   
## Min. :-1133   
## 1st Qu.:13184   
## Median :18838   
## Mean :19094   
## 3rd Qu.:24101   
## Max. :52062

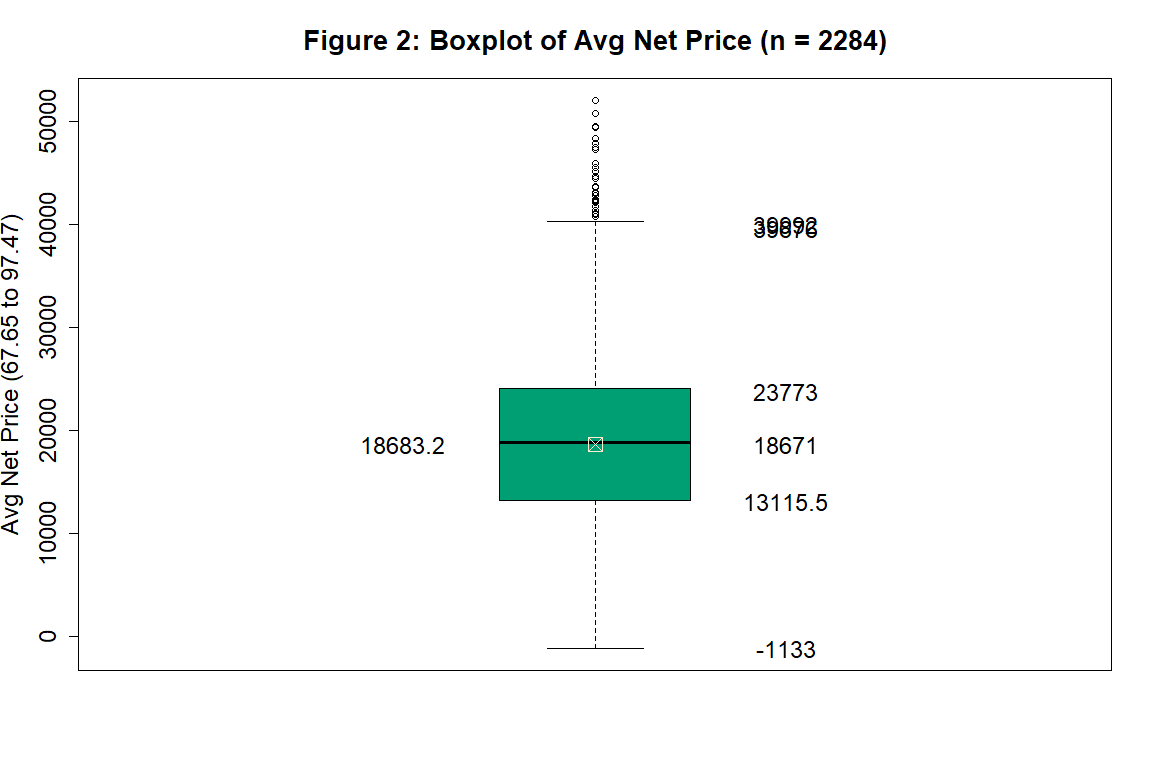
working\_TWO <- working %>% filter(NPT4\_ALL < 40000)  
  
working\_TWO %>% select(NPT4\_ALL) %>% summary()

## NPT4\_ALL   
## Min. :-1133   
## 1st Qu.:13116   
## Median :18671   
## Mean :18683   
## 3rd Qu.:23773   
## Max. :39992

vari <- working\_TWO$NPT4\_ALL  
  
boxplot(vari,  
 main="Figure 2: Boxplot of Avg Net Price (n = 2284)",  
 cex.main=1.7,  
 ylab="Avg Net Price (67.65 to 97.47)",  
 cex.lab=1.5,  
 boxwex=0.4,   
 cex.axis=1.5,  
 col= sample(cbPalette,1))   
text(y = boxplot.stats(vari)$stats,   
labels = round(boxplot.stats(vari)$stats,1), x = 1.20, cex = 1.5)  
text(y = round(max(vari),1), labels = round(max(vari),1), x = 1.20, cex = 1.5)  
points(mean(vari),pch=7, cex=1.9,col="blanchedalmond")   
text(y = round(mean(vari),1), labels = round(mean(vari),1), x = 0.80, cex = 1.5)



vari2 <- working$NPT4\_ALL  
  
boxplot(vari2,  
 main="Figure 2: Boxplot of Avg Net Price (n = 2284)",  
 cex.main=1.7,  
 ylab="Avg Net Price (67.65 to 97.47)",  
 cex.lab=1.5,  
 boxwex=0.4,   
 cex.axis=1.5,  
 col= sample(cbPalette,1))   
text(y = boxplot.stats(vari)$stats,   
labels = round(boxplot.stats(vari)$stats,1), x = 1.20, cex = 1.5)  
text(y = round(max(vari),1), labels = round(max(vari),1), x = 1.20, cex = 1.5)  
points(mean(vari),pch=7, cex=1.9,col="blanchedalmond")   
text(y = round(mean(vari),1), labels = round(mean(vari),1), x = 0.80, cex = 1.5)



#### **Background Information and Variable description**

      The data used to test the hypothesized value came from Department of Educations *Integrated Post-secondary Education Data System* (IPEDS). A copy of the raw data can be found at this [link](https://data.ed.gov/dataset/college-scorecard-all-data-files-through-6-2020/resources) by clicking the green button ‘Go To Resource’. One can also click this direct link [here](https://ed-public-download.app.cloud.gov/downloads/CollegeScorecard_Raw_Data_01192021.zip) to download the data directly from the College Scorecard API. This data was updated on December 20th, 2020 to include the 2017-2018 school year. The focus of this paper will be on the 2017-2018 school year data, in particular the variables corresponding to the average Net Price at an institution.

Variable Definitions

|  |  |  |  |
| --- | --- | --- | --- |
| *Variable* | *Definition* | *Type of institution* | *n* |
| *NPT4\_PUB* | Average net price for Title IV institutions. | Public institutions | 729 |
| *NPT4\_PRIV* | Average net price for Title IV institutions. | Private institutions | 1555 |

Hypothesis

|  |  |
| --- | --- |
| *Value:* | $18500 – |
|  | Average Net Price (in constant 2018-19 dollars) for (4-year) Title IV institutions. |
| *Claim:* | That the Average Net price at a (4-year) Title IV institution is $18500. |
| *Source:* | National Census for Educational Statistics (2017-18 school year). |
| *URL:* | <https://nces.ed.gov/programs/digest/d19/tables/dt19_331.30.asp> |

#### **Hypothesis and Results**

     The National Census for Educational Statistics (NCES) claims that the average Net Price at a (4-year) Title IV institution during the 2017-2018 school year is $18500. My data shows for the 2017-2018 school year the average Net Price is higher than the claimed value. In fact, we are 95% confident that the true average Net Price for (4-year) Title IV institutions is between $18800.2157 to $19387.4655. Some caution should be taken here; we have a large sample of 2284 (4-year) Title IV institutions and the data displays no evidence that suggests that the data could have came from a population that is normally distributed.  
      If a researcher is interested in testing if the median is equal to $18500, then the data does not contradict that the median is $18500. In fact, we are 95% confident that for the 2017-2018 school year the true median net price for a (4-year) Title IV institutions is between $18450 and $19265.  
      Since the data displayed no evidence that suggests that the sample came from a normally distributed population and our sample size is large (2284), the results from the nonparametric test on the median may yield better accuracy over the parametric results for the mean. This is because the results from the parametric test assumes that the mean is approximately close to the center of the distribution, whereas the results from the nonparametric test uses the sample’s distribution during the test on the median, and is therefore not affected by a skewed distribution.

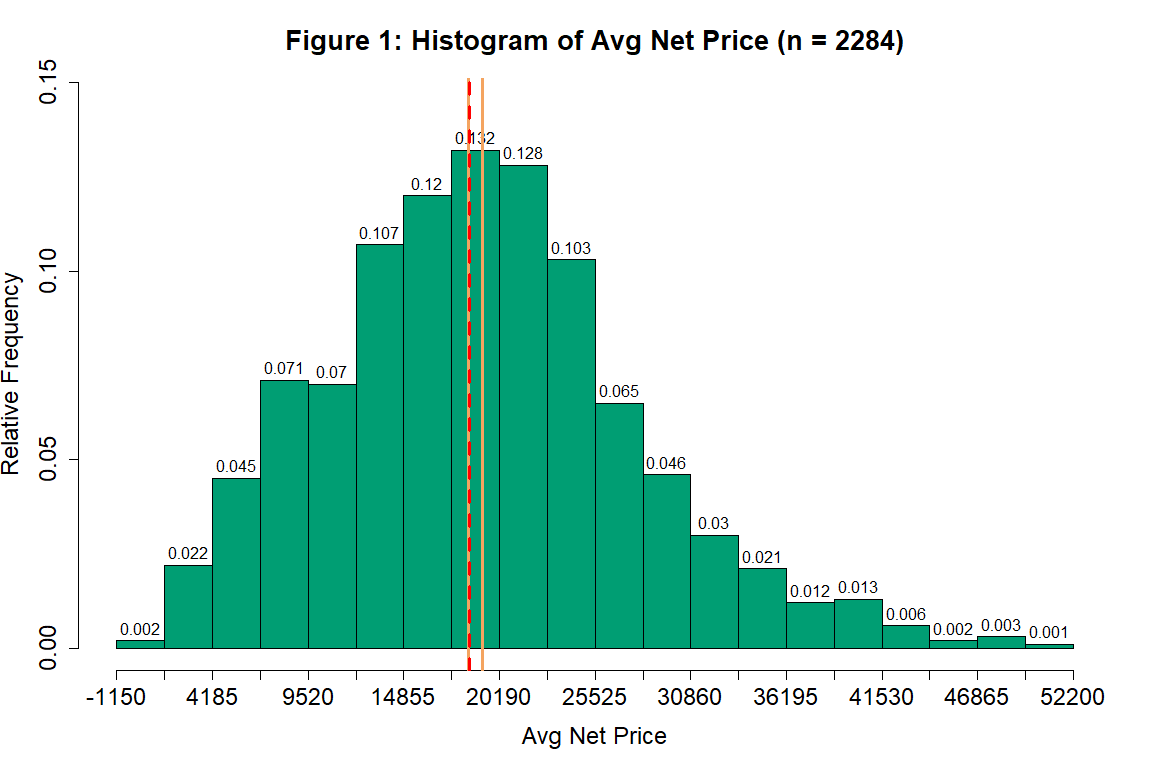
#### **Subset, Histogram, and Boxplot**

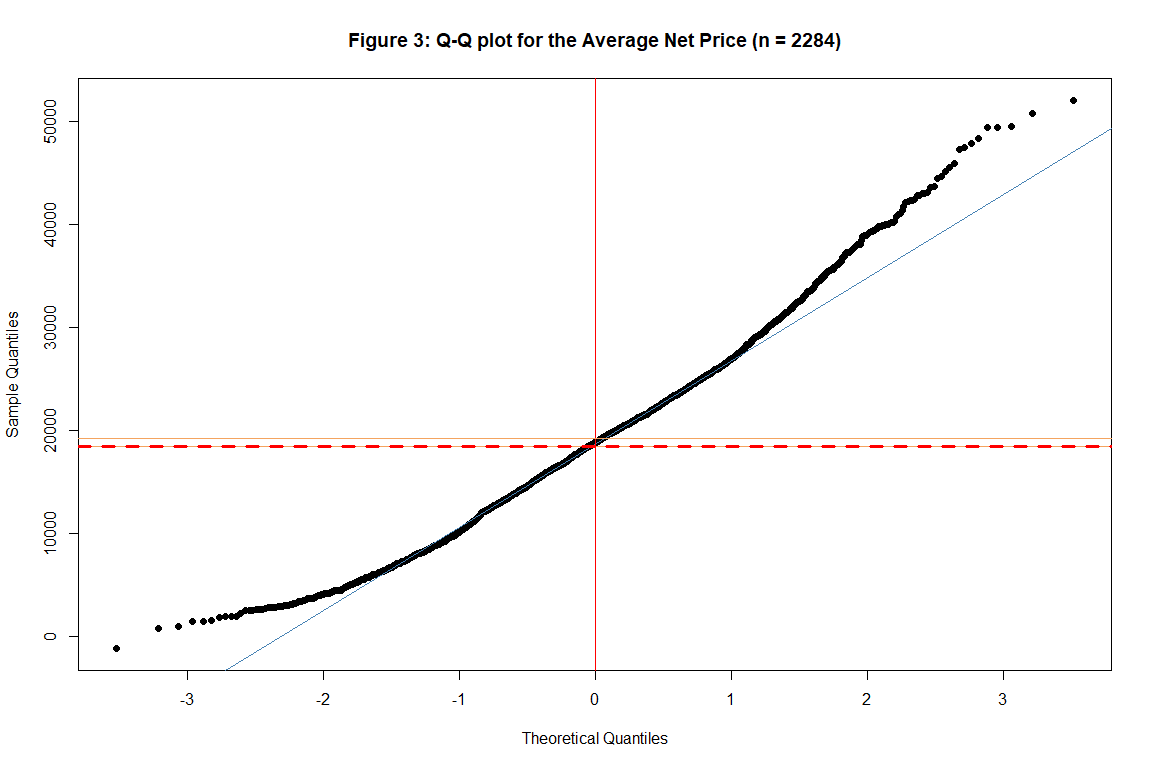
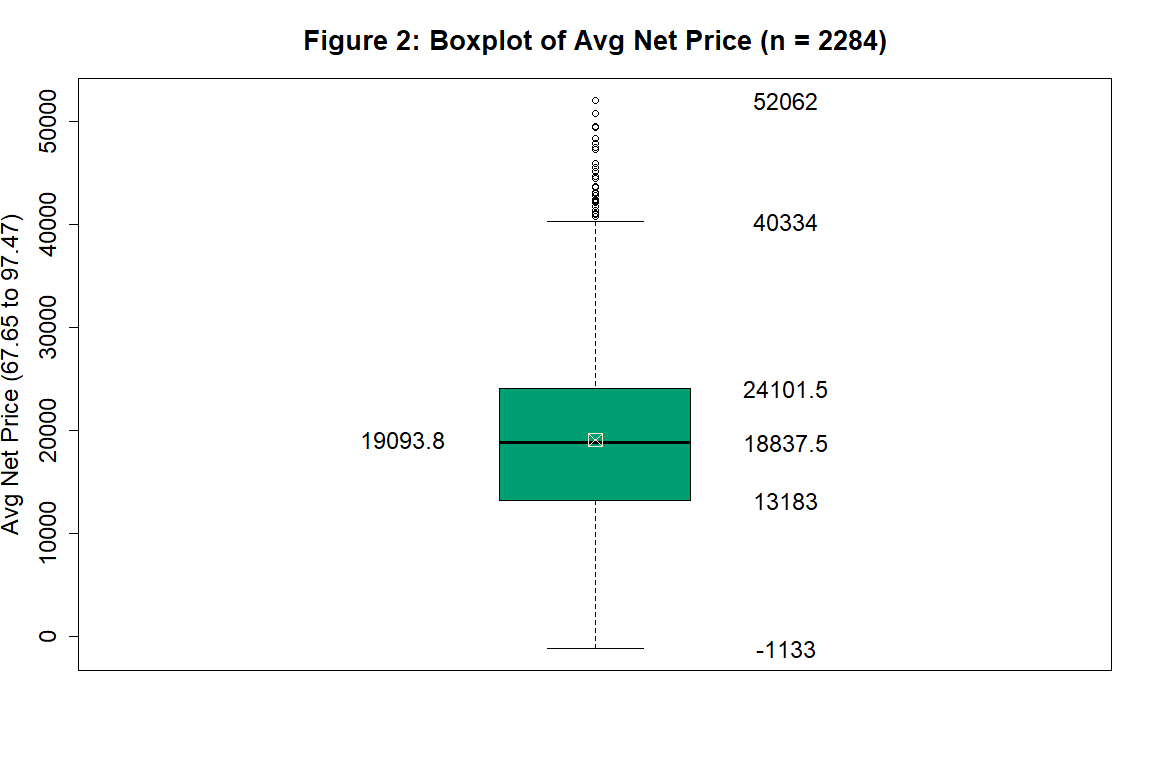
       For this hypothesis, I filtered in (4-year) Title IV institutions and merged the variables **NPT4\_PUB** and **NPT4\_PRIV** into one column. I labeled it **NPT4\_ALL**. Note that the variable prefix *NPT4\_* describes the average Net Price for all students at an institution. The suffix \*\_PUB\* associates the column with Public institutions while the suffix \*\_PRIV\* associates the column with Private institutions. Since my hypothesized value is for all (4-year) Title IV institutions, I merged these columns. The table below displays the descriptive statistics for the data when a Public (Left two columns) or a Private (Middle two columns) filter is applied. The rightmost columns display the descriptive statistics for the merged column containing both private and public institutions data.

**Descriptive Statistics for the Average Net Price (NPT4) at a (4-year) institution**    *Public Institutions data*              ##       *Private Institutions data*            ##       *Merged data*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| \*\*\_PUB | \*\*\_PRIV | ## | \*\*\_PUB | \*\*\_PRIV | ## | \*\*\_ALL |
| Min. :-1133 | Min. : NA | ## | Min. : NA | Min. : 1466 | ## | Min. :-1133 |
| Median :13293 | Median : NA | ## | Median : NA | Median :21571 | ## | Median :18838 |
| Mean :13010 | Mean :NaN | ## | Mean :NaN | Mean :21946 | ## | Mean :19094 |
| Max. :44661 | Max. : NA | ## | Max. : NA | Max. :52062 | ## | Max. :52062 |

      Next, a histogram, boxplot, and Q-Q plot were created from the column NPT4\_ALL. There are 2284 (4-year) Title IV institutions in the test sample. In twenty-five bins, the histogram shows a slight skewness to the right. To help visualize the results, three lines were placed on the histogram. The red dashed line is the hypothesized value of $18500. The two sandybrown lines correspond to the 95% confidence interval for the median ($18450, $19265). As one can see from the histogram in Figure 1, the values to the right of the confidence interval for the median become proportionally less frequent than values to the left. The boxplot in Figure 2 on the next page shows many outliers above the upper quartile, giving further evidence to suggest that the data is skewed right. From the Q-Q plot, the data shows a curve at each end of the line suggesting a non-linear relationship is present. On the vertical axis, three horizontal lines were placed corresponding to the hypothesized value( in red ) and the 95% confidence interval for the median( in sandybrown ). On the horizontal axis, a vertical line was placed in red at zero. These four lines partition the Q-Q plot into quadrants. From the Q-Q plot, one can see that the observations above the confidence interval’s upper bound ($19265) are distributed along a much larger range than observations below the confidence interval’s lower bound ($18450). With this information, we can conclude that there is not enough evidence to suggest that the data came from a normally distributed population.





#### **Shapiro-Wilk test**

The results of the Shapiro-Wilk normality test gave further evidence that the data is skewed. Since the p-value (8.81273406e-15) is less than the chosen level of confidence (5%), we can conclude that the data is not normally distributed.

#### **Conclusions from the parametric T-test**

       Since we received a p-value of (0.000444335174) and assumed a (5%) level of significance, we have enough evidence to reject the claimed mean of $18500. Therefore, we have enough evidence to conclude that there is a (0.0444335174)% chance of getting a random sample of 2284 institutions to have a mean Net Price of $18500 or less. This probability is sufficiently low enough to suggest that the true mean Net Price is higher than the hypothesized value of $18500. With this evidence, we can be 95% confident that the true mean Net Price for (4-year) Title IV institutions is between $18800.2157 to $19387.4655 for the 2017-2018 school year. As stated previously, since we do not have evidence to suggest that this data came from a normally distributed population, the results of this parametric test may yield a false-positive due to the data being skewed. The nonparametric test on the median should yield a more accurate results for this sample.

#### **Conclusions from the Nonparametric Binomial Distribution test**

       We received a p-value of 0.0582762273 which is slightly greater than the chosen level of confidence (0.05). This suggests that getting 1180 out of 2284 institutions having a average Net Price above the claimed value of $18500 is significantly more than the number of institutions expected to be above the median [(1142) = (2284)\*(0.5)]. With this information we have enough evidence to be 95.3191328% confident that the true median Net Price for a (4-year) Title IV institution is between $18450 and $19265 for the 2017-2018 school year. Considering that this interval contains the hypothesized value, we can conclude that the true median Net Price for a (4-year) Title IV institution can in fact be the claimed value ($18500).

#### **Conclusion**

      There is evidence to support the conclusion that the sample is skewed. Since the parametric T-test has less power to detect differences for skewed data, the results may yield a false-positive. The nonparametric binomial distribution test has more power to detect differences for skewed data. For this reason, it is suggested to use the results of the nonparametric test. That is, at a 95.3191328% level of confidence, we can conclude that the true median Net Price can in fact be the claimed value of $18500 since this value is in the 95.3191328% confidence interval. Further research with a more complex test, such as testing public and private institutions as their own category, may produce a more accurate conclusion.